Focused Active Inference

by

Daniel S. Levine

S.B., Massachusetts Institute of Technology (2008)
S.M., Massachusetts Institute of Technology (2010)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

[September 2014]
August 2014

Signature redacted

Author .....
Department of Aeronautics and Astronautics
August 2014

Signature redacted

Certified by .....
Jonathan P. How
Richard C. Maclaurin Professor of Aeronautics and Astronautics
Thesis Supervisor

Signature redacted

Certified by .....
John W. Fisher III
Senior Research Scientist, CSAIL

Signature redacted

Certified by .....
Nicholas Roy
Associate Professor of Aeronautics and Astronautics

Signature redacted

Accepted by .....
Paulo C. Lozano
Associate Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

Focused Active Inference

by

Daniel S. Levine

Submitted to the Department of Aeronautics and Astronautics
on August 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In resource-constrained inferential settings, uncertainty can be efficiently minimized with respect to a resource budget by incorporating the most informative subset of observations – a problem known as *active inference*. Yet despite the myriad recent advances in both understanding and streamlining inference through probabilistic graphical models, which represent the structural sparsity of distributions, the propagation of information measures in these graphs is less well understood. Furthermore, active inference is an **NP**-hard problem, thus motivating investigation of bounds on the suboptimality of heuristic observation selectors.

Prior work in active inference has considered only the *unfocused* problem, which assumes all latent states are of inferential interest. Often one learns a sparse, high-dimensional model from data and reuses that model for new queries that may arise. As any particular query involves only a subset of relevant latent states, this thesis explicitly considers the *focused* problem where irrelevant states are called *nuisance variables*. Marginalization of nuisances is potentially computationally expensive and induces a graph with less sparsity; observation selectors that treat nuisances as notionally relevant may fixate on reducing uncertainty in irrelevant dimensions. This thesis addresses two primary issues arising from the retention of nuisances in the problem and representing a gap in the existing observation selection literature.

The interposition of nuisances between observations and relevant latent states necessitates the derivation of *nonlocal* information measures. This thesis presents propagation algorithms for nonlocal mutual information (MI) on universally embedded paths in Gaussian graphical models, as well as algorithms for estimating MI on Gaussian graphs with cycles via embedded substructures, engendering a significant computational improvement over existing linear algebraic methods.

The presence of nuisances also undermines application of a technical diminishing returns condition called submodularity, which is typically used to bound the performance of greedy selection. This thesis introduces the concept of *submodular relaxations*, which can be used to generate online-computable performance bounds, and analyzes the class of optimal submodular relaxations providing the tightest such bounds.

Thesis Supervisor: Jonathan P. How
Title: Richard C. Maclaurin Professor of Aeronautics and Astronautics

Committee Member: John W. Fisher III
Title: Senior Research Scientist, CSAIL

Committee Member: Nicholas Roy
Title: Associate Professor of Aeronautics and Astronautics

Contents

List of Figures

Preface

Data are not costless; their acquisition or use requires the exchange of other resources, such as time or energy. Data are not equally useful; informativeness derives solely from the ability to resolve particular queries. Observations, realized as data, are like windows that permit partial glimpses into underlying states, whose values can never be exactly known and must be inferred. In assessing the informativeness of observations, it is not enough to consider the clarity of these windows; they must also face the right directions.

Chapter 1

Introduction

1.1 Problem Domain

This thesis addresses the problem of focused active inference: selecting a subset of observable random variables that is maximally informative with respect to a specified subset of latent random variables. The objective is to reduce the overall cost of inference while providing greater inferential performance. As only a portion of the potentially high-dimensional model is useful for resolving a given query, this thesis explicitly considers the common issue of nuisance variables, whose effective mitigation requires efficient algorithms for nonlocal information quantification and novel methods of bounding performance.

This section elucidates the three principal facets of the problem domain considered in this thesis: uncertainty reduction in Bayesian parameter estimation, analysis of informational relationships in probabilistic graphical models, and the inferential relevance of only a subset of latent variables constituent to high-dimensional models.

1.1.1 Active Parametric Uncertainty Reduction

In partially observable domains, latent states must be inferred from observations. The process that gives rise to realized observations, or *data*, rarely permits a deterministic mapping to the exact values of latent states. One must maintain a *belief*,

a probabilistic representation over possible configurations of latent states, with the eventual goal of reducing the uncertainty inherent in this belief.

Uncertainty reduction can be achieved through statistical inference, the melding of existing beliefs with statistics of data to form updated beliefs. However, acquiring or processing data requires an expense of resources (such as time, money, or energy), presenting a tradeoff between inferential performance and cost. Consideration of the *active inference* problem, in which a selector decides which observations to realize and supply to the inference engine, can lead to dramatic improvements in utilization of resources for uncertainty reduction. For example, control of the data acquisition process in sensor networks can lead to higher inferential performance under constraints on energy consumed via sensing, computation, and communication [16, 43].

The active inference problem can be specified with one of two "dual" formulations: minimizing cost subject to an information quota that must be achieved, or maximizing information subject to a cost constraint. This thesis, as in the larger literature of which it is a part (e.g., [34, 39, 63, 94]), prefers the latter formulation, as it is generally more natural to specify resource constraints than information quotas. (The quota formulation is further discussed in Section 6.2.3.) Likewise, specifying a "cost of ignorance" that scalarizes uncertainty in terms of a single resource — while providing a common metric between two fundamentally different quantities — may be very sensitive to the choice of the conversion factor.

The resource-constrained formulation can be solved by selecting observations based on a measure of informativeness. Provided that data are not deterministic, their expected informativeness is derived from the statistical model that relates their behavior to that of the underlying states one wishes to infer. This thesis considers informativeness of observations strictly in the context of Bayesian parameter estimation problems, the structure of which is described in the following subsection. (Other, more convoluted problem classes involving uncertainty reduction are discussed in Section 1.3.)

1.1.2 Graphical Model Analysis

This thesis assumes a model has been provided, and the ensuing goal is to interpret relationships within this model in the context of informativeness. This assumption is motivated by the hypothesis that, regardless of the specific sensing modalities or communication platforms used in an information collection system, the underlying phenomena can be described by *some* stochastic process structured according to a probabilistic graphical model. The sparsity of that model can potentially be exploited to yield efficient algorithms for inference [33, 37, 45]. It bears repeating that the model being analyzed is merely assumed. The relative appropriateness of the model is not asserted, as the upstream process of model selection is outside the scope of this thesis.

In the framework of this thesis, data sources are modeled as observable nodes in a probabilistic network. Subsets of informative data sources are then identified through algorithms developed in this thesis that, in analogy to graphical inference methods, capture how measures of informativeness propagate in graphs.

In contrast to methods for estimating information measures directly from raw data (e.g., [38]), the framework of this thesis: does not require the prior enumeration of interaction sets that one wishes to quantify; can capture relationships in models with latent variable structure [17, 54]; and can be used to compute *conditional* information measures that account for statistical redundancies between observations.

1.1.3 Focused Inference

It is often of benefit to learn a single, high-dimensional yet parsimonious graphical model (with potentially many latent variables) whose structure can be exploited for efficient inference [17, 37, 54, 85]. Given such a model, the objective of many inferential problems is to reduce uncertainty in only a *subset* of the unknown quantities. The set of *relevant* latent variables one wishes to infer may be respecified online as particular queries or applications arise. Irrelevant latent variables, or *nuisances*, are not of any extrinsic importance and act merely as intermediaries between observable variables and those of inferential interest.

One option is to simply marginalize out any nuisances, although this approach can be both computationally expensive and detrimental to the graph's sparsity, which, in the interest of efficient model utilization, one wishes to retain (cf. Section 2.3.1). As will soon be demonstrated (in Example 1 below), observation selectors that ignore the categoric distinctness of nuisance variables by treating them as notionally relevant can become fixated on reducing uncertainty in irrelevant portions of the underlying distribution. Therefore, this thesis proposes methods for selecting informative subsets of observations in graphical models that retain both nuisance variables in the joint distribution and their categoric distinctness in the selection process.

1.2 Problem Statement

Let $\mathbf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ be a collection of N random variables[1] indexed by $\mathcal{V} = \{1, \ldots, N\}$, and let the joint distribution $p_{\mathbf{x}}(\cdot)$ over $\mathbf{x}$ be Markov to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Consider a partition of $\mathcal{V}$ into the subsets of latent variables $\mathcal{U}$ and observable variables $\mathcal{S}$, with $\mathcal{R} \subseteq \mathcal{U}$ denoting the subset of *relevant* latent variables (i.e., those to be inferred). Given a cost function $c : 2^{\mathcal{S}} \to \mathbb{R}_{\geq 0}$ over subsets of observations and a budget $\beta \in \mathbb{R}_{\geq 0}$, the *focused* active inference problem is

$$\begin{aligned} \text{maximize}_{\mathcal{A} \subseteq \mathcal{S}} \quad & I(\mathbf{x}_{\mathcal{R}}; \mathbf{x}_{\mathcal{A}}) \\ \text{s.t.} \quad & c(\mathcal{A}) \leq \beta, \end{aligned} \tag{1.1}$$

where $I(\cdot\,;\cdot)$ is the mutual information measure (cf. Section 2.2.2).

[1] Throughout this thesis, random variables may be either scalar or vectoral. In the latter case, the elements of $(\mathsf{x}_1, \ldots, \mathsf{x}_N)$ are disjoint subvectors of $\mathbf{x}$.

1.2.1 Focused vs. Unfocused Selection

The focused active inference problem in (1.1) should be distinguished from the *unfocused* active inference problem typically considered (e.g., in [39, 94]),

$$\begin{aligned} \text{maximize}_{\mathcal{A}\subseteq\mathcal{S}} \quad & I(\mathbf{x}_{\mathcal{U}};\mathbf{x}_{\mathcal{A}}) \\ \text{s.t.} \quad & c(\mathcal{A}) \leq \beta, \end{aligned} \tag{1.2}$$

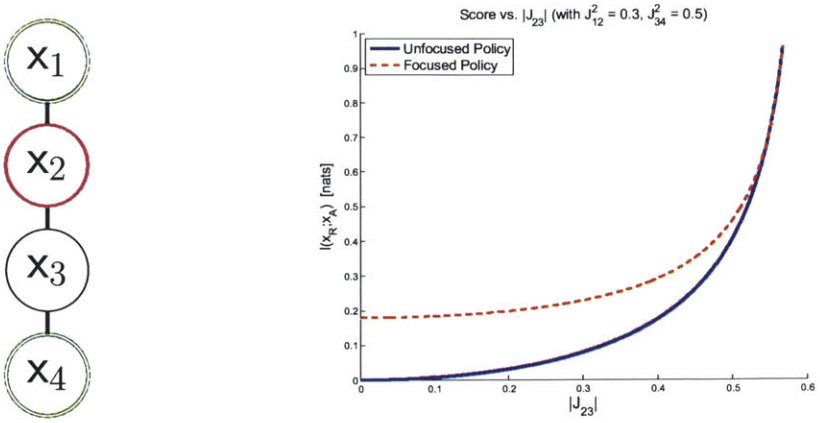
which regards the *entirety* of the latent state $\mathcal{U} \supseteq \mathcal{R}$ to be of interest. Both the focused and unfocused problems are known to be **NP**-hard [36, 41].

By the chain rule and nonnegativity of MI (reviewed in Section 2.2.2), for any $\mathcal{A} \subseteq \mathcal{S}$, $I(\mathbf{x}_{\mathcal{U}};\mathbf{x}_{\mathcal{A}}) = I(\mathbf{x}_{\mathcal{R}};\mathbf{x}_{\mathcal{A}}) + I(\mathbf{x}_{\mathcal{U}\setminus\mathcal{R}};\mathbf{x}_{\mathcal{A}}|\mathbf{x}_{\mathcal{R}}) \geq I(\mathbf{x}_{\mathcal{R}};\mathbf{x}_{\mathcal{A}})$. Therefore, maximizing the unfocused objective does not imply maximizing the focused objective. Focused active inference must be posed as a separate problem to preclude the possibility of the observation selector becoming fixated on inferring nuisance variables $\mathcal{U} \setminus \mathcal{R}$ as a result of $I(\mathbf{x}_{\mathcal{U}\setminus\mathcal{R}};\mathbf{x}_{\mathcal{A}} \mid \mathbf{x}_{\mathcal{R}})$ being implicitly included in the quantification. In fact, an unfocused selector can perform arbitrarily poorly with respect to a focused metric, as the following example illustrates.

Example 1. Let $\mathbf{x} = (\mathsf{x}_1, \mathsf{x}_2, \mathsf{x}_3, \mathsf{x}_4) \in \mathbb{R}^4$ be distributed according to a zero-mean Gaussian distribution $p_{\mathbf{x}}(\cdot) = \mathcal{N}(\cdot; \mathbf{0}, P)$ with symmetric, positive definite covariance

$$P = \begin{bmatrix} J_{11} & J_{12} & 0 & 0 \\ J_{21} & J_{22} & J_{23} & 0 \\ 0 & J_{32} & J_{33} & J_{34} \\ 0 & 0 & J_{43} & J_{44} \end{bmatrix}^{-1}, \tag{1.3}$$

where $J_{ij} = J_{ji} \in \mathbb{R}$. It can be shown (cf. Section 2.4.2) that $p_{\mathbf{x}}(\cdot)$ can be structured according to a four-vertex chain (Figure 1-1a). Suppose $\mathcal{R} = \{2\}$ is the relevant set, and that one observable variable from the set $\mathcal{S} = \{1, 4\}$ may be selected. The focused decision rule, which is strictly concerned with reducing uncertainty in x_2, has



(a) Graphical model. (b) Policy comparison.

Figure 1-1: (a) Graphical model for the four-node chain example. (b) Unfocused vs. focused policy comparison. The unfocused and focused policies coincide for large $|J_{23}|$; however, as $|J_{23}| \to 0^+$, the unfocused policy approaches complete performance loss with respect to the focused measure.

an optimal selection $\mathcal{A}^*(\mathcal{R}) = \text{argmax}_{a\in\{1,4\}} I(\mathsf{x}_2;\mathsf{x}_a)$ that can be shown to be [47]

$$|J_{23}| \cdot \mathbf{1}_{\{J_{34}^2 - J_{12}^2 J_{34}^2 - J_{12}^2 \geq 0\}} \underset{\mathcal{A}^*(\mathcal{R})=\{1\}}{\overset{\mathcal{A}^*(\mathcal{R})=\{4\}}{\gtrless}} \sqrt{\frac{(1-J_{34}^2)J_{12}^2}{J_{34}^2}},$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function, which evaluates to 1 when its argument is true and 0 otherwise.

Conversely, the *unfocused* decision rule, which considers the set of all latent states $\mathcal{U} = \{2,3\}$ to be of inferential interest, has an optimal selection $\mathcal{A}^*(\mathcal{U}) = \text{argmax}_{a\in\{1,4\}} I(\mathsf{x}_2,\mathsf{x}_3;\mathsf{x}_a)$ that can be reduced, by positive definiteness of P and through conditional independences in the chain graph, to

$$|J_{34}| \underset{\mathcal{A}^*(\mathcal{U})=\{1\}}{\overset{\mathcal{A}^*(\mathcal{U})=\{4\}}{\gtrless}} |J_{12}|.$$

The performance gap associated with optimizing the "wrong" information measure — i.e., using an unfocused selector for a focused uncertainty reduction problem — is demonstrated in Figure 1-1b. Note that the unfocused decision rule is independent of J_{23}, which effectively parameterizes the informational capacity of edge $\{2,3\}$. For

sufficiently large $|J_{23}|$, the two policies coincide and achieve the same performance. However, as $|J_{23}| \to 0^+$, the information that x_3 can convey about x_2 also approaches zero, although the unfocused decision rule is oblivious to this fact.

1.3 Related Research

The concept of informative planning for uncertainty reduction has a long and varied history, subsuming the not entirely overlapping fields of design of experiment [5, 25, 28, 51], informative motion planning [7, 46, 49, 58, 72], active learning [18, 26, 36, 56, 62, 77, 78, 87, 88], sensor management [12, 43, 75, 93], adaptive sampling [15, 24, 58, 66], and belief space planning [9, 70, 73, 76, 89]. The area of probabilistic graphical models [33, 37, 45] has developed more recently and is arguably more cohesive. However, despite abundant interest in probabilistic graphical models and informative planning, the intersection of these two sets is far from voluminous.

One possible explanation for this gap is that the informative planning literature has been primarily application-oriented. Sensing platforms and the external phenomena that they are intended to sense are often modeled separately, without a common computational structure on which to compute belief updates. The resulting optimization over observation selections is necessarily myopic, as the inability to exploit structure in the underlying distributions can complexify the quantification of long-horizon informativeness. This would not be as problematic if the reward structure had a local decomposition. However, in many estimation problems (e.g., simultaneous localization and mapping [31, 79]), uncertainty reduction may be driven by events that unfold over long horizons in spacetime. In the focused problems of interest in this thesis, myopic observation strategies might be uniformly uninformative about the states of inferential interest.

This section reviews the existing research most germane to, and in the context of, the focused active inference problem of (1.1).

1.3.1 Focused Quantification and Goal-Oriented Inference

The focused problem class explored in this thesis is most similar to the adaptive sampling problem considered in [15], wherein a sensor network selects an observation plan in spacetime that maximizes mutual information with respect to numerical weather predictions located in a *subset* of spatial states at some terminal time. The framework of this thesis can be seen as a generalization of [15] insofar as the relevant set is an arbitrary subset of latent states and is not constrained to lie in a single slice of spacetime. The informativeness quantifications of [15] were computed with ensemble numerical methods and did not exploit sparsity via a graphical model representation.

There are also parallels between focused quantification and the goal-oriented inference framework of Lieberman for identifying or approximating sufficient parameter subspaces in PDE-constrained stochastic systems where only a subset of prediction dimensions are relevant [50], although the problem domains of the cited paper and this thesis are, at least presently, disjoint.

1.3.2 Information-theoretic and Abstract Rewards

Uncertainty reduction is a critical capability for statistical artificial intelligence [86]. In partially observable sequential decision processes, stochasticities (e.g., noisy transition and emission dynamics) can induce uncertainty in the reward under a particular policy. One method, referred to as *belief space planning*, involves the construction of a notional augmented state space in which each configuration is a belief, i.e., a distribution over configurations of a lower-dimensional underlying state [8, 35]. Uncertainty reduction actions may be executed as part of a heuristic that trades off information-gathering and reward seeking (exploration vs. exploitation) [11, 52]. In the partially observable Markov decision process (POMDP) framework, an abstract reward function can be convolved with the transition and emission distributions, inducing for each policy a distribution over the return from any initial belief state. However, the same generality that makes this framework appealing also leads to serious issues of tractability that effectively limit the degree to which global uncertainty can be effi-

ciently minimized. Despite this limitation, planners with local abstract rewards have found abundant use. In robot navigation problems, for example, active mitigation of uncertainty may be required to prevent the robot's state estimate from diverging [9, 73].

Active inference can be seen as a specific case of belief space planning with a purely information-theoretic reward function [42]. One can make the following analogy: active inference is to general belief space planning (i.e., with rewards not exclusive to uncertainty reduction) as estimation is to direct adaptive control. In estimation, the performance measure is exclusive to statistics on the error signal of the state estimates. In direct adaptive control, estimation performance is not specified and is merely a byproduct of attempting to improve control performance. One consequence is that it may be problematic to extract a measure of plan informativeness from variations in abstract reward functions. This thesis considers purely information-theoretic rewards, which may be nonlocal due to nuisances. A broader goal is the clarification of uncertainty reduction properties that would enable better scalability for belief space planning with general, abstract rewards.

1.3.3 Active Learning and Informativeness

An arguably orthogonal problem to active inference is *active learning* [18, 26, 36], wherein a set of unlabeled, partially labeled, or incomplete examples are used to learn a model, which may include structure and/or parameters. (For a comprehensive review of active learning, see [78].) As in active inference, the overall goal of active learning is to reduce labeling/completion querying costs while improving the quality of the model.

Active learning has typically been attempted to reduce the overall labeling cost in semisupervised discriminative model learning settings. For example, Lizotte et al. actively learn a naïve Bayes *discriminative* model under a constrained querying budget; the problem is cast as a Markov decision process (MDP), and heuristic solutions akin to the greedy algorithm are presented [56].

There are, however, examples of active learning of generative models. Tong and

Koller consider two problem types that involve actively learning a Bayesian network (directed model) [87, 88] with fewer labeled examples than needed when learning from independent, random samples of the underlying distribution. In [87], Tong and Koller actively learn the parameters of a hierarchical categorical distribution (i.e., over a finite alphabet) with known structure. The algorithm proposed in [87] is based on a greedy reduction in the expected posterior risk resulting from making a query to the network that returns a labeled example. The manipulation of the risk calculation based on properties of the Dirichlet distribution — the conjugate prior to the categorical distribution — affords tractability. (Tong and Koller also consider actively learning the structure of a Bayesian network [88].)

Anderson and Moore present several objective functions, and subsequently algorithms, for actively learning models, states, and paths through homogeneous hidden Markov models (HMMs) [3]. They later extended their results to minimizing user-specified risk functions in directed polytrees; all vertices are assumed to be selectable for observing, and a comparison between their proposed method and mutual information quantification assumes an unfocused setting [4].

1.3.4 Active Inference Performance Bounds

The intersection of graphical model analysis and observation selection has been explored most extensively in the recent active inference literature, primarily to derive bounds on the suboptimality of heuristic observation selectors in the unfocused setting [27, 39, 94]. These bounds are all based on variations of the greedy selection heuristic and differ depending on whether the bounds: hold a priori or are computed online; are open-loop or closed-loop with respect to realizations of selected observations; or assume homogeneous or heterogeneous costs for observations. The structure inherent in all such bounds developed thus far has assumed *unfocused* selection.

Krause and Guestrin leveraged seminal work on *submodularity* [64] to bound the suboptimality of open-loop, greedy unfocused selections of unit-cost observations [39]. Similar bounds have also been established when the costs are heterogeneous [40]. Williams et al. provide bounds when the underlying state has additional latent struc-

ture and when there exist selection constraints [94]. The analysis in [94] also leads to tighter online-computable suboptimality bounds — using the set of next-best observable variables *not* selected by the greedy algorithm at each stage — and diffusive bounds based on local information measures in graph neighborhoods. A generalization of submodularity called *adaptive submodularity* has also been introduced to derive closed-loop bounds when the realized observations may be incorporated into future observation selection decisions [27].

1.4 Main Issues

The contributions of this thesis are directed towards resolving the two primary issues that arise when retaining — i.e, neither marginalizing out nor willfully ignoring — nuisance variables in the problem: nonlocality and nonsubmodularity of the information measure. Both issues are unresolved in the existing literature and represent gaps bridged in this thesis.

1.4.1 Nonlocality

In the unfocused problem (1.2) typically considered in the literature, since all latent variables are of inferential interest, the information measures that must be quantified are local to the Markov blanket of every observation [39, 94]. In the focused problem (1.1), observable and relevant latent variables may be nonadjacent in the graphical model due to the interposition of nuisances between them. Thus, focused observation selection requires the development of information measures that extend beyond adjacency, or *locality*, in the graph. To be of practical use, such nonlocal information measures must also be efficiently computable.

1.4.2 Nonsubmodularity

Given the exponential complexity inherent in verifying the optimality of solutions to both the focused and unfocused problems, practitioners of active inference require

methods for bounding the suboptimality of heuristically generated solutions. In the unfocused problem, under very mild additional assumptions about the Markov structure of the graph, performance bounds derived from submodularity can be invoked [39, 64, 94]. In the focused problem, the absence of certain conditional independences, particularly between observations when conditioned on the relevant latent variable set, violates the assumptions that permit application of existing submodularity-based performance bounds [47]. Thus, focused active inference requires the development of new methods for bounding performance of heuristic selections.

1.5 Summary of Thesis Contributions

Chapter 2 reviews preliminary results, primarily in graphical inference and information theory, upon which the contributions of this thesis are constructed. Those contributions include:

- Decomposition of nonlocal mutual information on universally embedded paths in scalar and vectoral Gaussian graphical models, enabling a reduction in the complexity of greedy selection updates on Gaussian trees from cubic to linear in the network size [Chapter 3, first published in [47]];
- Efficient information quantifications in loopy Gaussian graphs by leveraging the computational structure in embedded tree estimation of conditional covariance matrices [Chapter 4, first published in [48]];
- Introduction of *submodular relaxations* in general (i.e., non-Gaussian) undirected graphical models to derive new, online-computable suboptimality bounds for observation selection in the presence of nuisance variables, which invalidate previous submodular bounds; and characterization of the optimal submodular relaxation providing the tightest such bounds, as well as heuristics for approximating the optimum [Chapter 5].

Chapter 2

Preliminaries

This chapter presents general background material upon which the contributions of this thesis in subsequent chapters are predicated. Additional material specific to individual chapters will be presented where appropriate.

2.1 Bayesian Inference

Let $p_{\mathbf{x}}(\cdot)$ denote the *joint distribution* over a collection $\mathbf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ of N random variables (or disjoint random subvectors) indexed by the set $\mathcal{V} = \{1, \ldots, N\}$. For some subset $A \subset \mathcal{V}$, the *marginal distribution* $p_{\mathbf{x}_A}(\cdot)$ over A may be calculated by marginalizing out the components of $\mathbf{x}$ indexed by $B := \mathcal{V} \setminus A$ according to

$$p_{\mathbf{x}_A}(\cdot) = \int_{\mathcal{X}_B} p_{\mathbf{x}}(\mathbf{x}_A, \mathbf{x}_B)\, d\mathbf{x}_B, \tag{2.1}$$

where $\mathcal{X}_B = \prod_{b \in B} \mathcal{X}_b$ is the space of all feasible joint configurations of $\mathbf{x}_B$.

For disjoint $A, C \subset \mathcal{V}$, the *posterior distribution* $p_{\mathbf{x}_A|\mathbf{x}_C}(\cdot|\mathbf{x}_C)$ over $\mathbf{x}_A$ given the realization $\mathbf{x}_C := \mathbf{x}_C$ may be calculated via Bayes rule, i.e.,

$$p_{\mathbf{x}_A|\mathbf{x}_C}(\mathbf{x}_A|\mathbf{x}_C) = \frac{p_{\mathbf{x}_A,\mathbf{x}_C}(\mathbf{x}_A, \mathbf{x}_C)}{\int_{\mathcal{X}_A} p_{\mathbf{x}_A,\mathbf{x}_C}(\mathbf{x}'_A, \mathbf{x}_C)\, d\mathbf{x}'_A}. \tag{2.2}$$

2.2 Information-theoretic Measures

This section reviews the key results that lead up to the definition, and some useful properties, of mutual information (MI), the information measure central to this thesis. MI is arguably a fundamental statistical quantity, but other measures have been used to actively select observations. The entropy measure indicates *where* there is uncertainty but, as has been noted, does not capture the degree to which it can be reduced [74]. A synopsis of the "alphabetical" optimality criteria of experimental design [5, 51] and their relationships with active inference can be found in [42]. Ali-Silvey information measures, also known as f-divergences, can be viewed as generalizations of MI and are discussed as future work in Section 6.2.6.

2.2.1 Entropy

Entropy is one measure of the uncertainty entailed by a probabilistic distribution. As this thesis primarily considers real-valued data sources, a review of Shannon entropy for discrete random variables is provided merely to build intuition. The rest of this subsection proceeds with the differential form for continuous random variables.

Shannon Entropy

For a discrete random vector $\mathbf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ with alphabet $\mathcal{X}_\mathcal{V}$ and probability mass function (pmf) $p_\mathbf{x}(\cdot)$, the Shannon entropy $H(\mathbf{x})$ is defined as [19]

$$H(\mathbf{x}) \triangleq \mathbb{E}_{p_\mathbf{x}}\left[\log \frac{1}{p_\mathbf{x}(\mathbf{x})}\right] = -\sum_{\mathbf{x}\in\mathcal{X}_\mathcal{V}} p_\mathbf{x}(\mathbf{x}) \log p_\mathbf{x}(\mathbf{x}), \tag{2.3}$$

where the base of the logarithm determines the units (e.g., base e for nats, and base 2 for bits). As $0 \leq p_\mathbf{x}(\mathbf{x}) \leq 1$ for all $\mathbf{x} \in \mathcal{X}_\mathcal{V}$ (by the definition of a pmf), then $H(\mathbf{x}) \geq 0$. One interpretation of Shannon entropy is the average length of the shortest description of a random variable, e.g., in bits. If $\mathcal{X}_\mathcal{V}$ is finite, the maximum-entropy pmf over $\mathbf{x}$

can be shown to be the uniform distribution

$$p_{\mathbf{x}}^{ME}(\mathbf{x}) = \texttt{uniform}(\mathbf{x}; \mathcal{X}_{\mathcal{V}}) \triangleq \begin{cases} \frac{1}{|\mathcal{X}_{\mathcal{V}}|}, & \mathbf{x} \in \mathcal{X}_{\mathcal{V}}, \\ 0, & \text{otherwise}, \end{cases} \tag{2.4}$$

whereby $0 \leq H(\mathbf{x}) \leq \log |\mathcal{X}_{\mathcal{V}}|$.

Differential Entropy

For a continuous random vector $\mathbf{x}$ with probability density function (pdf) $p_{\mathbf{x}}(\cdot)$ over the support set $\mathcal{X}_{\mathcal{V}}$, the differential entropy $\mathcal{H}(\mathbf{x})$ is defined as

$$\mathcal{H}(\mathbf{x}) \triangleq \mathbb{E}_{p_{\mathbf{x}}} \left[\log \frac{1}{p_{\mathbf{x}}(\mathbf{x})} \right] = - \int_{\mathcal{X}_{\mathcal{V}}} p_{\mathbf{x}}(\mathbf{x}) \log p_{\mathbf{x}}(\mathbf{x}) \, d\mathbf{x}. \tag{2.5}$$

Note that $p_{\mathbf{x}}(\cdot) \geq 0$ but is not generally bounded from above at unity, so differential entropy is not always nonnegative.[1]

If the partition $\mathcal{V} = A \dot{\cup} B$ (with $\dot{\cup}$ denoting the union of disjoint sets) induces marginal distributions $p_{\mathbf{x}_A}(\cdot)$ and $p_{\mathbf{x}_B}(\cdot)$, then the conditional differential entropy $\mathcal{H}(\mathbf{x}_A|\mathbf{x}_B)$ of $\mathbf{x}_A$ given $\mathbf{x}_B$ is

$$\mathcal{H}(\mathbf{x}_A|\mathbf{x}_B) = - \int p_{\mathbf{x}}(\mathbf{x}_A, \mathbf{x}_B) \log p_{\mathbf{x}_A|\mathbf{x}_B}(\mathbf{x}_A|\mathbf{x}_B) \, d\mathbf{x}_A \, d\mathbf{x}_B \tag{2.6}$$

$$= \mathcal{H}(\mathbf{x}_A, \mathbf{x}_B) - \mathcal{H}(\mathbf{x}_B). \tag{2.7}$$

Relative Entropy

Let $\mathbf{x} \in \mathcal{X}_{\mathcal{V}}$ be a continuous random vector, and let p and q be two densities over $\mathbf{x}$. The relative entropy, or Kullback-Leibler divergence, between p and q is

$$D(p \parallel q) \triangleq \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \, d\mathbf{x}. \tag{2.8}$$

[1]This somewhat limits intuitive interpretations of differential entropy. One must also take care in handling differential entropies that are infinite.

Note that relative entropy is not generally symmetric, i.e., $D(p \parallel q) \neq D(q \parallel p)$, and is finite only if the support set of p is contained in that of q. (By convention and as motivated by continuity, let $0 \log \frac{0}{0} := 0 =: 0 \log 0$.) One particularly notable result, due to Jensen's inequality, is that

$$D(p \parallel q) \geq 0, \tag{2.9}$$

with equality if and only if $p = q$ almost everywhere. Therefore, even though differential entropy can be negative, relative entropy is always nonnegative.

2.2.2 Mutual Information

Mutual information (MI) is an information-theoretic measure of dependence between two (sets of) random variables. Its interpretation as a measure of entropy reduction appeals to its use in uncertainty mitigation [10].

Let $\mathsf{x} \in \mathcal{X}$ and $\mathsf{y} \in \mathcal{Y}$ be two random variables with joint distribution $p_{\mathsf{x},\mathsf{y}}$ and marginal distributions p_{x} and p_{y}. Mutual information is a particular evaluation of relative entropy, namely,

$$I(\mathsf{x};\mathsf{y}) \triangleq D(p_{\mathsf{x},\mathsf{y}} \parallel p_{\mathsf{x}} p_{\mathsf{y}}) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_{\mathsf{x},\mathsf{y}}(x,y) \log \frac{p_{\mathsf{x},\mathsf{y}}(x,y)}{p_{\mathsf{x}}(x) p_{\mathsf{y}}(y)} \, dy \, dx. \tag{2.10}$$

Because relative entropy is always nonnegative, so is MI. Some manipulation of the MI definition immediately leads to

$$\begin{aligned} I(\mathsf{x};\mathsf{y}) &= \mathcal{H}(\mathsf{x}) - \mathcal{H}(\mathsf{x}|\mathsf{y}) \\ &= \mathcal{H}(\mathsf{x}) + \mathcal{H}(\mathsf{y}) - \mathcal{H}(\mathsf{x},\mathsf{y}) \\ &= \mathcal{H}(\mathsf{y}) - \mathcal{H}(\mathsf{y}|\mathsf{x}) \\ &= I(\mathsf{y};\mathsf{x}), \end{aligned}$$

so MI is symmetric with respect to its (nonconditional) arguments.

Of particular concern in this thesis are conditional MI measures. Consider the

continuous random vector $\mathbf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N) \in \mathcal{X}_\mathcal{V}$, with $\mathcal{V} = \{1, \ldots, N\}$. For disjoint subsets $A, B, C \subset \mathcal{V}$ (with C possibly empty), conditional mutual information can be specified by augmenting all entropy terms with the appropriate conditioning set, i.e.,

$$I(\mathbf{x}_A; \mathbf{x}_B | \mathbf{x}_C) \triangleq \mathcal{H}(\mathbf{x}_A | \mathbf{x}_C) + \mathcal{H}(\mathbf{x}_B | \mathbf{x}_C) - \mathcal{H}(\mathbf{x}_{A \cup B} | \mathbf{x}_C). \tag{2.11}$$

For convenience, the index sets will often be used as arguments of mutual information in place of the random variables they index, i.e., $I(A; B|C) \triangleq I(\mathbf{x}_A; \mathbf{x}_B | \mathbf{x}_C)$. Manipulation of entropy terms leads to the chain rule of MI,

$$I(A; B \cup C) = I(A; C) + I(A; B|C). \tag{2.12}$$

From Equation (2.9), it is clear that

$$I(A; B|C) \geq 0, \quad \text{with equality iff } \mathbf{x}_A \perp\!\!\!\perp \mathbf{x}_B \mid \mathbf{x}_C, \tag{2.13}$$

where the last statement may be parsed as, "$\mathbf{x}_A$ is conditionally independent of $\mathbf{x}_B$ when conditioned on $\mathbf{x}_C$." An intuitive result of the above is the "information never hurts" principle,

$$\mathcal{H}(A|B \cup C) \leq \mathcal{H}(A|C), \quad \text{with equality iff } \mathbf{x}_A \perp\!\!\!\perp \mathbf{x}_B \mid \mathbf{x}_C. \tag{2.14}$$

2.3 Probabilistic Graphical Models

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$ linking pairs of vertices, can be used to represent the conditional independence structure of a joint distribution $p_\mathbf{x}(\cdot)$ over a collection $\mathbf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ of N random variables (or disjoint random subvectors). While there exist several classes of probabilistic graphical models [37, 45], many inferential procedures are currently best understood for undirected graphs. Alternative models are briefly discussed in Section 2.3.2.

2.3.1 Undirected Graphs

Graph Topology

An undirected graphical model, or Markov random field (MRF), has an undirected edge set $\mathcal{E} \subseteq \binom{\mathcal{V}}{2}$, where $\binom{\mathcal{V}}{2}$ denotes the set of all unordered pairs of distinct vertices drawn from $\mathcal{V}$. The neighborhood function $\Gamma : 2^{\mathcal{V}} \to 2^{\mathcal{V}}$ returns the vertices adjacent to any input vertex subset $A \subseteq \mathcal{V}$ and is defined such that $\Gamma(A) \triangleq \{j : \exists \{i, j\} \in \mathcal{E} \text{ for some } i \in A\}$.

The topology of an undirected graph $\mathcal{G}$ can be characterized, in part, by its set of paths. A path is a sequence of distinct adjacent vertices $(v_1, \ldots, v_m)$ where $\{v_k, v_{k+1}\} \in \mathcal{E}$, $k = 1, \ldots, m-1$. The diameter of $\mathcal{G}$, denoted as $\text{diam}(\mathcal{G})$, is the length of its longest path.

For disjoint subsets $A, B, C \subset \mathcal{V}$, where C is possibly empty, let $\mathcal{P}(A, B|C)$ denote the set of all paths between every $u \in A$ and $v \in B$ in $\mathcal{G}$ that do not include vertices in C.[2] If $\mathcal{P}(A, B|C) = \emptyset$, then A and B are *graph disconnected* conditioned on C. If $|\mathcal{P}(u, v)| = 1$ for distinct $u, v \in \mathcal{V}$, then there is a *unique* path between u and v, and the sole element of $\mathcal{P}(u, v)$ is denoted by $\bar{\pi}_{u:v}$. If $|\mathcal{P}(u, v)| > 1$ for some $u, v \in \mathcal{P}$, the $\mathcal{G}$ contains a cycle and is called "loopy."

A graph without cycles is called a *tree* (or, if disconnected, a disjoint union of trees appropriately called a *forest*). A *chain* is a simple tree with $\text{diam}(\mathcal{G}) = |\mathcal{V}|$. A potentially loopy graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ contains a number of useful embedded substructures. A graph $\mathcal{G}_\mathcal{T} = (\mathcal{V}_\mathcal{T}, \mathcal{E}_\mathcal{T})$ without cycles is a spanning tree of $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if $\mathcal{V}_\mathcal{T} = \mathcal{V}$ and $\mathcal{E}_\mathcal{T} \subseteq \mathcal{E}$. A maximal spanning tree $\mathcal{G}_\mathcal{T}$ of $\mathcal{G}$ cannot be further augmented with an edge from $\mathcal{E} \setminus \mathcal{E}_\mathcal{T}$ without introducing a cycle. A path is *universally embedded* in $\mathcal{G}$ if it is contained in every maximal spanning tree embedded in $\mathcal{G}$.

[2] Alternatively, one could specify paths in the subgraph induced by $\mathcal{V} \setminus C$. Since this thesis is framed with the problem of reusing a single high-dimensional model, and since conditioning in undirected graphs entails simply "deafening" vertices in the conditioning set, directly specifying the conditioning set wherever possible is preferred over induced subgraph notation.

Conditional Independence

An MRF can represent conditional independences of the form given by the global Markov condition: $\mathbf{x}_A \perp\!\!\!\perp \mathbf{x}_B \mid \mathbf{x}_C$ iff $\mathcal{P}(A, B|C) = \emptyset$. A distribution is said to be Markov with respect to a graph $\mathcal{G}$ if it satisfies the conditional independences implied by $\mathcal{G}$.

Let $CI(\cdot)$ denote the set of conditional independences either satisfied by a distribution or implied by a graph. If $CI(\mathcal{G}) \subseteq CI(p)$, then $\mathcal{G}$ is an independence map, or *I-map*, of p. The trivial I-map is a fully connected graph. A minimal I-map is one such that adding any further conditional independences renders it no longer an I-map; a minimal *undirected* I-map is one such that no edge may be removed without violating the I-map condition. Any particular distribution has within the set of all of its undirected I-maps a single *minimal* undirected I-map [37]. It is assumed throughout this thesis that the structure of the distribution is represented by its unique minimal undirected I-map.

Conversely, if $CI(p) \subseteq CI(\mathcal{G})$, then $\mathcal{G}$ is a dependence map, or *D-map*, of p. The trivial D-map is a fully disconnected graph, which has no edges. D-maps can be used as embedded substructures to perform iterative belief updates, as illustrated in Chapter 4.

Factorization

A subset $C \subset \mathcal{V}$ is a clique in $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if for all distinct $i, j \in C$, $\{i, j\} \in \mathcal{E}$. A clique C is maximal if no strict superset $C' \supset C$ is also a clique. The Hammersley-Clifford theorem [30] relates the structure of a graph to the factorization of all distributions that satisfy its implied conditional independence structure: If $p_{\mathbf{x}}(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}$, then for any undirected graph $\mathcal{G}$ to which p is Markov,

$$p_{\mathbf{x}}(\mathbf{x}) \propto \prod_{C \in \mathcal{C}} \psi_C(\mathbf{x}_C), \tag{2.15}$$

where $\mathcal{C}$ is the set of all maximal cliques of $\mathcal{G}$, and for each $C \in \mathcal{C}$, $\psi_C(\cdot)$ is a nonnegative *potential function* that assigns a weight to every joint configuration $\mathbf{x}_C$

Algorithm 1 StructuralElimination($\mathcal{V}$, $\mathcal{E}$, $B \subset \mathcal{V}$, $(b_1, \ldots b_{|B|})$)

Require: undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $B \subset \mathcal{V}$, elimination ordering $(b_1, \ldots, b_{|B|})$

$\tilde{\mathcal{V}} \leftarrow \mathcal{V}$, $\tilde{\mathcal{E}} \leftarrow \mathcal{E}$, $\tilde{\mathcal{G}} \leftarrow (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$

for $k = 1, \ldots, |B|$ **do**

$\tilde{\mathcal{V}} \leftarrow \tilde{\mathcal{V}} \setminus \{b_k\}$

$\tilde{\mathcal{E}} \leftarrow \tilde{\mathcal{E}} \cup \big\{\{i,j\} : i,j \in \Gamma_{\tilde{\mathcal{G}}}(b_k),\ i \neq j\big\}$

$\tilde{\mathcal{E}} \leftarrow \tilde{\mathcal{E}} \setminus \{\{i,j\} \in \mathcal{E} : i = b_k\}$

$\tilde{\mathcal{G}} \leftarrow (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$

end for

return $\tilde{\mathcal{G}}$

of random subvectors $\mathbf{x}_C$ participating in C.

Structural Elimination

Let disjoint subsets $A, B, C \subset \mathcal{V}$ form a partition of $\mathcal{V}$, and let $\mathbf{b} = (b_1, \ldots, b_{|B|})$ be an ordering over the elements of B. Marginalizing out $\mathbf{x}_B$ from the joint distribution under elimination order $\mathbf{b}$ involves the integration

$$p_{\mathbf{x}_A, \mathbf{x}_C}(\mathbf{x}_A, \mathbf{x}_C) = \int_{x_{b_{|B|}} \in \mathcal{X}} \cdots \int_{x_{b_1} \in \mathcal{X}} p(\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C)\, dx_{b_1} \ldots\, dx_{b_{|B|}}. \tag{2.16}$$

By substituting the factorization (2.15), it is straightforward to show that the integrations of (2.16) induce intermediate marginal distributions Markov to *marginal graphs* $\tilde{\mathcal{G}}_1, \ldots, \tilde{\mathcal{G}}_{|B|}$. Consider the innermost integration, which corresponds to marginalizing out b_1. Cliques in which b_1 does not participate may be moved outside this integral. The resulting definite integral generates a new potential function corresponding to a maximal clique over $\Gamma(b_1)$.

Given only a graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the structural elimination algorithm (Algorithm 1) can be used to examine the marginal graphs resulting from eliminating random variables indexed by $B \subset \mathcal{V}$ under a particular elimination ordering $(b_1, \ldots, b_{|B|})$. Consider eliminating b_k, $k \in \{2, \ldots, |B|\}$. The resulting graph $\tilde{\mathcal{G}}_k = (\tilde{\mathcal{V}}_k, \tilde{\mathcal{E}}_k)$ will have a vertex set $\tilde{\mathcal{V}}_k = \tilde{\mathcal{V}}_{k-1} \setminus \{b_k\}$ and edge set $\tilde{\mathcal{E}}_k = \tilde{\mathcal{E}}_{k-1} \cup \Big\{\{i,j\} : i,j \in \Gamma_{\tilde{\mathcal{G}}_{k-1}}(b_k), i \neq j\Big\} \setminus \Big\{\{i,j\} \in \tilde{\mathcal{E}}_{k-1} : i = b_k\Big\}$.

An example of structural elimination is depicted in Figure 2-1, from which it is

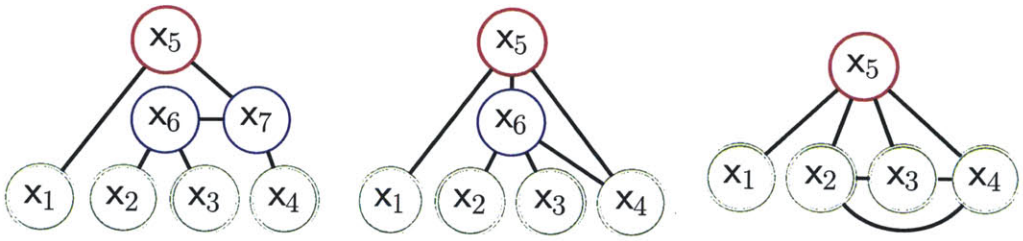


Figure 2-1: Example of marginalizing out nuisances under the elimination order $(7, 6)$.

clear that marginalization of nuisances, in addition to being potentially computationally expensive (as incurred by (2.16)), mars the sparsity of the graph.

2.3.2 Alternative models

Several graphical model classes provide an alternative to undirected graphs. Factor graphs [57] provide a more fine-grained factorization of the distribution than is permissible in (2.15). Directed models [68] factorize the joint distribution into conditional probability distributions along directed edges in an acyclic graph and can be used to imply causal relationships. The sets of conditional independences that can be implied by directed and undirected models are generally not completely overlapping. The extension of the results of this thesis to factor and directed graphs is an area of future work.

2.4 Multivariate Gaussian Distributions

This section examines the special case in which $\mathbf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ is distributed according to a (nondegenerate) multivariate Gaussian

$$p_{\mathbf{x}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu, P) \triangleq \frac{1}{|2\pi P|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T P^{-1}(\mathbf{x}-\mu)\right\} \tag{2.17}$$

with mean vector $\mu \triangleq \mathbb{E}[\mathbf{x}]$ and (symmetric, positive definite) covariance matrix $P \triangleq \mathbb{E}\left[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\right]$. Assume without loss of generality[3] that each component

[3]The extension to varying subvector dimensions with $d \triangleq \max_{i \in \mathcal{V}} \dim(\mathsf{x}_i)$ is straightforward.

x_i of $\mathbf{x}$ is a subvector of dimension $d \in \mathbb{N}^+$, whereby $P \in \mathbb{R}^{Nd \times Nd}$ can be partitioned into an $N \times N$ grid of $d \times d$ block submatrices.

A multivariate Gaussian distribution can alternatively be represented in the information form $p_{\mathbf{x}}(\mathbf{x}) = \mathcal{N}^{-1}(\mathbf{x}; \mathbf{h}, J) \propto \exp\{-\frac{1}{2}\mathbf{x}^T J\mathbf{x} + \mathbf{h}^T\mathbf{x}\}$, with (symmetric, positive definite) *precision* or *inverse covariance matrix* $J = P^{-1}$ and potential vector $\mathbf{h} = J\mu$. Estimating the mean of a Gaussian is then equivalent to solving the system of equations

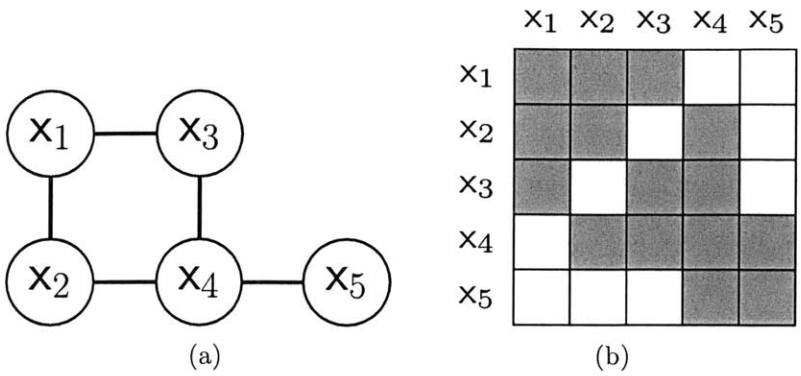
$$J\widehat{\mathbf{x}} = \mathbf{h}. \tag{2.18}$$

2.4.1 Fundamental Properties

Gaussian distributions have a number of fundamental properties that motivate the study of information measure propagation in this thesis. The Gaussian is the only absolutely continuous distribution that can be completely described by its first and second moments; it is also the maximum entropy distribution for a given mean and covariance. It is often used to approximate real-valued unimodal data or, due to the central limit theorem (and under weak conditions), the sum of many arbitrarily distributed random variables. Gaussians are also closed under various operations, such as marginalization and conditioning, the latter of which importantly results in a conditional covariance whose value is independent of the *realized configuration* of the conditioning random variable.

2.4.2 Gaussian MRFs

If $\mathbf{x} \sim \mathcal{N}^{-1}(\mathbf{h}, J)$, the conditional independence structure of $p_{\mathbf{x}}(\cdot)$ can be represented with a Gaussian MRF (GMRF) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E}$ is determined by the sparsity pattern of J and the pairwise Markov property: $J_{ij} = (J_{ji})^T \neq \mathbf{0}$ if and only if $\{i, j\} \in \mathcal{E}$ [82]. (An example of an undirected graph $\mathcal{G}$ and the sparsity pattern of a precision matrix J Markov to $\mathcal{G}$ is given in Figure 2-2.) Conversely, the covariance matrix P is generally not sparse and does not satisfy any useful global Markov properties.



(a) (b)

Figure 2-2: (a) An example undirected graph $\mathcal{G}$. (b) Sparsity pattern of precision matrix J Markov to $\mathcal{G}$. The structure of the graph can be read directly from $\mathcal{G}$ or from the nonzero elements of J.

In a scalar ($d = 1$) GMRF, $\mathcal{V}$ indexes scalar components of $\mathbf{x}$. In a *vectoral* ($d \geq 1$) GMRF, $\mathcal{V}$ indexes disjoint subvectors of $\mathbf{x}$, and the block submatrix J_{ii} can be thought of as specifying the sparsity pattern of the scalar micro-network within the vectoral macro-node $i \in \mathcal{V}$. Thus, it is possible to have a loopy graph over Nd scalar vertices that can be partitioned into a tree-shaped graph over N vertices of dimension d, which can be referred to as a vectoral tree.

2.4.3 Marginalization and Conditioning

Marginalization and conditioning can be conceptualized as selecting submatrices of P and J, respectively. Of course, in the inferential setting, one has access to J and not P.

Let disjoint sets A and B form a partition of $\mathcal{V} = \{1, \ldots, N\}$, so that

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{pmatrix} \sim \mathcal{N}^{-1}\left(\begin{pmatrix} \mathbf{h}_A \\ \mathbf{h}_B \end{pmatrix}, \begin{bmatrix} J_{AA} & J_{AB} \\ J_{AB}^T & J_{BB} \end{bmatrix}\right)$$
$$= \mathcal{N}\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{bmatrix} P_{AA} & P_{AB} \\ P_{AB}^T & P_{BB} \end{bmatrix}\right)$$

The marginal distribution $p_{\mathbf{x}_A}(\cdot)$ over $\mathbf{x}_A$ is parameterized by mean vector μ_A and covariance matrix $P_A = P_{AA}$, the subvector of μ and block submatrix of P corresponding

to the rows and columns indexed by A. Alternatively, the marginal distribution in information form $p_{\mathbf{x}_A}(\cdot) = \mathcal{N}^{-1}(\cdot; \widehat{\mathbf{h}}_A, \widehat{J}_A)$ is parameterized by $\widehat{\mathbf{h}}_A = \mathbf{h}_A - J_{AB}J_{BB}^{-1}\mathbf{h}_B$ and $\widehat{J}_A = J_{AA} - J_{AB}J_{BB}^{-1}J_{AB}^T$, the latter being the Schur complement of J_{BB}.

Conditioning on a particular realization $\mathbf{x}_B$ of the random subvector $\mathbf{x}_B$ induces the conditional distribution $p_{\mathbf{x}_A|\mathbf{x}_B}(\cdot|\mathbf{x}_B) = \mathcal{N}^{-1}(\cdot\,; \widehat{\mathbf{h}}_{A|B}, \widehat{J}_{A|B})$, where $\widehat{\mathbf{h}}_{A|B} = \mathbf{h}_A - J_{AB}\mathbf{x}_B$, and $\widehat{J}_{A|B} = J_{AA}$ is exactly the upper-left block submatrix of J. It is especially noteworthy that the conditional precision matrix is *independent of the value* of the realized configuration $\mathbf{x}_B$.

2.4.4 Gaussian Belief Propagation (GaBP)

Solving (2.18) by inverting J requires $\mathcal{O}((Nd)^3)$ operations, which can be prohibitively expensive for large N. If the graph contains no cycles, then Gaussian belief propagation (GaBP) [68, 91] can be used to compute the conditional mean, as well as marginal variances, in $\mathcal{O}(Nd^3)$, providing a significant computational savings for large N. For graphs with cycles, various estimation procedures have been recently developed to exploit available sparsity in the graph (cf. Sections 4.1.1 and 4.1.2).

Belief propagation (BP) [68] algorithms are variational inference procedures that use message passing on a graphical structure to approximate local posterior beliefs. BP is appealing due, in part, to its graphically intuitive mechanics, its amenability to implementation and parallelization, and the degree to which it generalizes many existing algorithms across disparate fields (e.g., [61]).

For Gaussian distributions, starting from the initial message set $J_{i\to j}^{(0)} := \mathbf{0}$ and $\mathbf{h}_{i\to j}^{(0)} := \mathbf{0}$ for all $\{i,j\} \in \mathcal{E}$, the GaBP message passing update equations at time $t+1$ are

$$\begin{aligned} J_{i\to j}^{(t+1)} &= -J_{ji}\left(J_{ii} + \sum_{k\in\Gamma(i)\setminus\{j\}} J_{k\to i}^{(t)}\right)^{-1} J_{ij} \\ \mathbf{h}_{i\to j}^{(t+1)} &= -J_{ji}\left(J_{ii} + \sum_{k\in\Gamma(i)\setminus\{j\}} J_{k\to i}^{(t)}\right)^{-1}\left(\mathbf{h}_i + \sum_{k\in\Gamma(i)\setminus j} \mathbf{h}_{k\to i}^{(t)}\right). \end{aligned} \tag{2.19}$$

For all $i \in \mathcal{V}$, marginal distributions $p_{\mathbf{x}_i}(\cdot) = \mathcal{N}^{-1}(\cdot; \widehat{\mathbf{h}}_{\{i\}}, \widehat{J}_{\{i\}})$ for all $i \in \mathcal{V}$ can

then be determined by using messages incident to i to compute the marginal precision matrix $\widehat{J}_{\{i\}}$ and marginal potential vector $\widehat{\mathbf{h}}_{\{i\}}$ according to

$$\begin{aligned} \widehat{J}_{\{i\}} &= J_{ii} + \sum_{k \in \Gamma(i)} J_{k \to i} \\ \widehat{\mathbf{h}}_{\{i\}} &= \mathbf{h}_i + \sum_{k \in \Gamma(i)} \mathbf{h}_{k \to i}, \end{aligned} \tag{2.20}$$

from which the marginal covariances $P_i = \left(\widehat{J}_{\{i\}}\right)^{-1}$ and marginal means $\mu_i = P_i \widehat{\mathbf{h}}_{\{i\}}$, for all $i \in \mathcal{V}$, can be computed in $\mathcal{O}(Nd^3)$.

Conditioning on $\mathbf{x}_C$, $C \subset \mathcal{V}$, can be handled in a variety of ways. Perhaps the most convenient method is to "deafen" conditioned nodes by prohibiting them from receiving or processing exogenous messages; if necessary (e.g., for conditional mean estimation), they can still author their own messages.

For a tree-shaped Gaussian MRF $\mathcal{G}$, it is also possible to compute other local beliefs. Given a pair of adjacent vertices $i, j \in \mathcal{V}$ such that $\{i, j\} \in \mathcal{E}$, the joint edge-marginal distribution $p_{\mathbf{x}_i, \mathbf{x}_j}(\cdot, \cdot)$ can be computed by forming the edge-marginal precision matrix and potential vector

$$\widehat{J}_{\{i,j\}} = \begin{bmatrix} \widehat{J}_{i \setminus j} & J_{ij} \\ J_{ji} & \widehat{J}_{j \setminus i} \end{bmatrix}, \qquad \widehat{\mathbf{h}}_{\{i,j\}} = \begin{bmatrix} \widehat{\mathbf{h}}_{i \setminus j} \\ \widehat{\mathbf{h}}_{j \setminus i} \end{bmatrix}, \tag{2.21}$$

where

$$\widehat{J}_{i \setminus j} = J_{ii} + \sum_{k \in \Gamma(i) \setminus \{j\}} J_{k \to i}, \qquad \widehat{\mathbf{h}}_{i \setminus j} = \mathbf{h}_i + \sum_{k \in \Gamma(i) \setminus \{j\}} \mathbf{h}_{l \to j},$$

with symmetric corrections for the other terms.

Parallel GaBP

The message passing updates of (2.19) can be implemented in parallel, which is of benefit for high-dimensional distributions. For Gaussian trees, parallel GaBP is guaranteed to converge to the exact vertex- and edge-marginal distributions in $\mathrm{diam}(\mathcal{G})$ iterations. The overall complexity of parallel GaBP is then $\mathcal{O}\left(\mathrm{diam}(\mathcal{G}) \cdot Nd^3\right)$: at

each of the diam($\mathcal{G}$) iterations until convergence, all N nodes perform message updates, each of which requires $\mathcal{O}(d^3)$ operations to compute.

Parallel GaBP on loopy Gaussian graphs (dubbed *loopy GaBP*) can be viewed as a variational approximation that attempts to minimize Bethe free energy, where the variational distribution is "tree-like" [95]. In general, loopy GaBP, when it does converge, does so to the correct conditional mean but to incorrect vertex-marginal variances [91]. There are subclasses of loopy GMRFs (such as diagonally dominant and walk-summable models [59]) for which convergence in both means and variances is guaranteed. Other graphical inference techniques for loopy Gaussian graphs are discussed in Sections 4.1.1 and 4.1.2.

Serial GaBP

For Gaussian trees, the GaBP updates can also be implemented in serial by noting that vertices at different tiers of the graph may cease to update their messages before iteration $t = \text{diam}(\mathcal{G})$. As $\mathcal{G}$ is an undirected graph, consider an arbitrarily chosen "root" $r \in \mathcal{V}$. All nodes $i \in \mathcal{V} \setminus \{r\}$ such that $|\Gamma(i)| = 1$ are called "leaves." Serial GaBP proceeds by having the leaf vertices update their messages, then all neighbors of leaves, and so on up to the root node. When the root node has received all its incoming messages, it begins disseminating messages back to its neighbors, down the tree again, to the leaf vertices. At this point all vertices in the tree will have a complete set of incident messages. The complexity of this approach is $\mathcal{O}(Nd^3)$: there are $\mathcal{O}(N)$ sequential message updates, each one requiring $\mathcal{O}(d^3)$ operations to form.

Covariance analysis

The graphical inference community appears to best understand the convergence of message passing algorithms for continuous distributions on subclasses of multivariate Gaussians (e.g., tree-shaped [91], walk-summable [59], and feedback-separable [53] models, among others). Gaussian trees comprise an important class of distributions that subsumes Gaussian hidden Markov models (HMMs), and GaBP on trees is a generalization of the Kalman filtering/smoothing algorithms that operate on HMMs.

Just as covariance analysis is used to analyze the evolution of uncertainty in the standard Kalman filter [60, 80], one of the goals of this thesis is to provide tools for performing general *preposterior analysis*, first for Gaussian trees (cf. Chapter 3), then for the class of general Gaussian distributions (cf. Chapter 4).

2.4.5 Gaussian Information Measures

If $\mathbf{x} \sim \mathcal{N}(\mu, P)$, where $\mu \in \mathbb{R}^n$ and $P \in \mathbb{R}^{n \times n}$, then the (differential) entropy of $\mathbf{x}$ is [19]

$$\mathcal{H}(\mathbf{x}) = \frac{1}{2} \log \left((2\pi e)^n \cdot \det(P) \right), \tag{2.22}$$

where e is the natural base. For a given subset $A \subset \mathcal{V}$, the marginal entropy of subvector $\mathbf{x}_A$ can be computed by replacing the full covariance matrix in (2.22) with P_A, the covariance parameterizing the marginal distribution.

Let $A, B \subset \mathcal{V}$ be disjoint, and consider the conditional distribution $p_{\mathbf{x}_A|\mathbf{x}_B}(\cdot|\mathbf{x}_B)$ of subvector $\mathbf{x}_A$ conditioned on a realization $\mathbf{x}_B = \mathbf{x}_B$. Since this distribution is only over $\mathbf{x}_A \in \mathbb{R}^d$, application of Equation (2.22) implies

$$\mathcal{H}(\mathbf{x}_A|\mathbf{x}_B = \mathbf{x}_B) = \frac{1}{2} \log \left((2\pi e)^d \cdot \det(P_{A|B}) \right), \tag{2.23}$$

where $P_{A|B}$ is the covariance matrix parameterizing $p_{\mathbf{x}_A|\mathbf{x}_B}(\cdot|\mathbf{x}_B)$, independent of the actual realized value $\mathbf{x}_B$ (cf. Section 2.4.3). Thus, the conditional entropy of $\mathbf{x}_A$ given $\mathbf{x}_B$ is

$$\begin{aligned} \mathcal{H}(\mathbf{x}_A|\mathbf{x}_B) &= \mathbb{E}_{\mathbf{x}_B} \left[\mathcal{H}(\mathbf{x}_A|\mathbf{x}_B = \mathbf{x}_B) \right] \\ &= \frac{1}{2} \log \left((2\pi e)^d \cdot \det(P_{A|B}) \right). \end{aligned} \tag{2.24}$$

Let $P_{A|C}$ denote the covariance of $\mathbf{x}_A$ given $\mathbf{x}_C$. For multivariate Gaussians, the

conditional MI [19] is

$$\begin{aligned} I(A;B|C) &= \mathcal{H}(\mathbf{x}_A|\mathbf{x}_C) + \mathcal{H}(\mathbf{x}_B|\mathbf{x}_C) - \mathcal{H}(\mathbf{x}_A, \mathbf{x}_B|\mathbf{x}_C) \\ &= \frac{1}{2}\log\frac{\det(P_{A|C})\det(P_{B|C})}{\det(P_{A\cup B|C})}. \end{aligned} \tag{2.25}$$

The most important implication of (2.24) and (2.25) is that Gaussian mutual information is *value independent* in the sense that it is in no way influenced by the actual realized configurations of observable random variables. This suggests that in Gaussian domains, there is no difference between forming an open-loop information collection plan (all observations selected before any is realized) and a closed-loop policy (selecting an observation, realizing it, processing it, and selecting the next observation accordingly).

Computing the marginal covariance matrices needed in (2.25) via matrix inversion (or taking Schur complements) of $J_{A\cup B|C}$ generally requires $\mathcal{O}((Nd)^3)$ operations, even if one is computing *pairwise* MI (i.e., $|A| = |B| = 1$). However, in light of Equations (2.21) and (2.25), pairwise MI quantities between *adjacent* vertices $i, j \in \mathcal{V}$ may be expressed as

$$I(i;j|C) = \frac{1}{2}\log\det(P_{i|C}) + \frac{1}{2}\log\det(P_{j|C}) - \frac{1}{2}\log\det(P_{\{i,j\}|C}), \quad \{i,j\} \in \mathcal{E}, \tag{2.26}$$

i.e., purely in terms of vertex- and edge-marginal covariance matrices. Thus, for Gaussian trees, GaBP provides a way of computing all *local* pairwise MI quantities in $\mathcal{O}(Nd^3)$. The decomposition of *nonlocal* MI into the sum of transformed local MI terms is the subject of Chapter 3.

2.5 Greedy Selection

One heuristic commonly used in combinatorial optimization problems is greedy selection, which is appealing due its computational simplicity and its relationship to the optimality of matroidal problems [67]. In general, the greedy heuristic is suboptimal,

Algorithm 2 GREEDYSELECTION-MAX(Z, f, c, β)

Require: ground set Z, objective function $f : 2^Z \to \mathbb{R}$, subadditive cost function $c : 2^Z \to \mathbb{R}_{\geq 0}$, budget $\beta \geq 0$

$\mathcal{A} \leftarrow \emptyset$
$Z_{\text{feas}} \leftarrow \{z \in Z : c(z) \leq \beta\}$
while $Z_{\text{feas}} \neq \emptyset$ **do**
  $\bar{a} \leftarrow \operatorname{argmax}_{a \in Z_{\text{feas}}} f(\mathcal{A} \cup \{a\})$
  $\mathcal{A} \leftarrow \mathcal{A} \cup \{\bar{a}\}$
  $Z_{\text{feas}} \leftarrow \{z \in Z \setminus \mathcal{A} : c(z) \leq \beta - c(\mathcal{A})\}$
end while
return $\mathcal{A}_\beta^g := \mathcal{A}$

although its suboptimality can in some cases be bounded; alternatively, it can be used to establish (maximization) upper bounds or (minimization) lower bounds on the optimal objective.

Given a ground set Z, an objective function $f : 2^Z \to \mathbb{R}$ one wishes to maximize, a subadditive cost function $c : 2^Z \to \mathbb{R}_{\geq 0}$ (i.e., such that $c(\mathcal{A}) \leq \sum_{a \in \mathcal{A}} c(a)$ for all $\mathcal{A} \in 2^Z$), and a budget $\beta \geq 0$, the general greedy selection template in Algorithm 2 sequentially selects the element with the highest marginal gain until the budget is expended (or until no further selections are affordable).

The greedy heuristic will be referenced throughout this thesis. In Chapter 5, a technical diminishing returns property called submodularity will be reviewed in the context of existing unfocused performance bounds and subsequently extended to the case of distributions with nuisances. The special case of subset selection that will be examined is unit-cost observations ($c(\mathcal{A}) = |\mathcal{A}|$ for all $\mathcal{A} \in 2^Z$) and bounded-cardinality constraints (corresponding to $\beta \in \mathbb{N}$). Bounds for unfocused selection with heterogeneous costs are considered in [40].

Chapter 3

Gaussian Tree Information Quantification

In an effort to pave the way for analyzing focused active inference on a broader class of distributions, this chapter specifically examines multivariate Gaussian distributions — which exhibit a number of properties amenable to analysis — that are Markov to tree shaped graphs. This chapter presents a decomposition of pairwise nonlocal mutual information (MI) measures on universally embedded paths in Gaussian graphs that permits efficient information valuation, e.g., to be used in a greedy selection. Both the valuation and subsequent selection may be distributed over nodes in the network, which can be of benefit for high-dimensional distributions and/or large-scale distributed sensor networks.

3.1 Information Efficiency

One of the underlying hypotheses of this chapter is that on a universally embedded path in a graph, the pairwise mutual information between its endpoints should fall off monotonically. This hypothesis is based on the understanding that graphical models represent statistical dependencies, and when two adjacent nodes are free to take on "different" configurations (i.e., when relationships are not deterministic), information should diminish accordingly on the edge connecting them in the graph.

In analogy to physical systems, consider the concept of efficiency. The presence of dissipative forces implies that any efficiency η must satisfy $\eta \in [0, 1]$, as in the first law of thermodynamics. In trying to derive work, for example, there are generally many constituent energy conversion efficiencies that must be multiplied to obtain the overall efficiency. The hypothesis of this chapter is that informational relationships can be specified with similar efficiencies: For any chain of scalar vertices that is universally embedded in a graph, the (nonlocal) mutual information between its endpoints is not only monotonically decreasing with the length of the chain, it can be computed as the product of *locally* defined efficiencies. The decompositions of nonlocal MI for Gaussian trees, as presented in the following section, makes explicit the intuitions of this hypothesis.

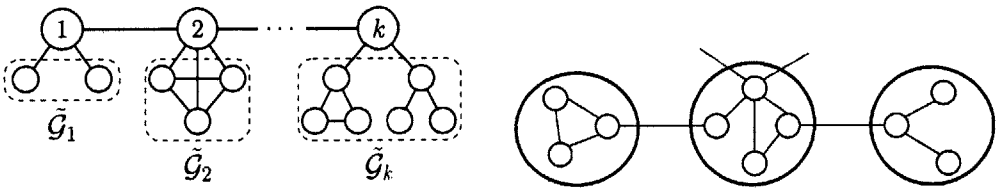
3.2 Nonlocal MI Decomposition

For GMRFs with N nodes indexing d-dimensional random subvectors, $I(\mathbf{x}_{\mathcal{R}}; \mathbf{x}_{\mathcal{A}})$ can be computed exactly in $\mathcal{O}((Nd)^3)$ via Schur complements/inversions on the precision matrix J. However, certain graph structures permit the computation via belief propagation of all local pairwise MI terms $I(\mathsf{x}_i; \mathsf{x}_j)$, for adjacent nodes $i, j \in \mathcal{V}$ in $\mathcal{O}(Nd^3)$ — a substantial savings for large networks. This section describes a transformation of nonlocal MI between uniquely path-connected vertices that permits a decomposition into the sum of transformed local MI quantities, i.e., those relating adjacent nodes in the graph. Furthermore, the local MI terms can be transformed in constant time, yielding an $\mathcal{O}(Nd^3)$ algorithm for computing any pairwise nonlocal MI quantity coinciding with a universally embedded path.

Definition 1 (Warped MI). For disjoint subsets $A, B, C \subseteq \mathcal{V}$, the warped mutual information measure $W : 2^{\mathcal{V}} \times 2^{\mathcal{V}} \times 2^{\mathcal{V}} \to (-\infty, 0]$ is defined such that $W(A; B|C) \triangleq \frac{1}{2} \log \eta(A; B|C)$, where $\eta(A; B|C) \triangleq (1 - \exp\{-2I(\mathbf{x}_A; \mathbf{x}_B|\mathbf{x}_C)\}) \in [0, 1]$.

For convenience, let $W(i; j|C) \triangleq W(\{i\}; \{j\}|C)$ for $i, j \in \mathcal{V}$.

Remark 2. For $i, j \in \mathcal{V}$ indexing scalar nodes, the warped MI of Definition 1 reduces to $W(i; j) = \log |\rho_{ij}|$, where $\rho_{ij} \in [-1, 1]$ is the correlation coefficient between scalar



(a) Unique path with sidegraphs. (b) Vectoral graph with thin edges.

Figure 3-1: (a) Example of a nontree graph $\mathcal{G}$ with a unique path $\bar{\pi}_{1:k}$ between nodes 1 and k. The "sidegraph" attached to each node $i \in \bar{\pi}_{1:k}$ is labeled as $\tilde{\mathcal{G}}_i$. (b) Example of a vectoral graph with thin edges, with internal (scalar) structure depicted.

r.v.s x_i and x_j. The measure $\log|\rho_{ij}|$ has long been known to the graphical model learning community as an "additive tree distance" [17, 23], and the decomposition for vectoral graphs presented here is a novel application for sensor selection problems. To the best of the author's knowledge, the only other distribution class with established additive distances are tree-shaped symmetric discrete distributions [17], which require a very limiting parameterization of the potentials functions defined over edges in the factorization of the joint distribution.

Proposition 3 (Scalar Nonlocal MI Decomposition). *For any GMRF $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ indexes scalar random variables, if $|\mathcal{P}(u, v|C)| = 1$ for distinct vertices $u, v \in \mathcal{V}$ and $C \subseteq \mathcal{V} \setminus \{u, v\}$, then $I(\mathsf{x}_u; \mathsf{x}_v|\mathsf{x}_C)$ can be decomposed as*

$$W(u; v|C) = \sum_{\{i,j\} \in \mathcal{E}(\bar{\pi}_{u:v|C})} W(i; j|C), \tag{3.1}$$

where $\mathcal{E}(\bar{\pi}_{u:v|C})$ is the set of edges joining consecutive vertices of $\bar{\pi}_{u:v|C}$, the sole element of $\mathcal{P}(u, v|C)$.

(Proofs of this and subsequent propositions can be found in Section 3.6.)

Remark 4. Proposition 3 requires only that the conditional path between vertices u and v be unique. If $\mathcal{G}$ is a tree, this is obviously satisfied. However, the result holds for any induced subgraph $\mathcal{G}(\mathcal{V} \setminus C)$ in which the path between u and v is embedded in *every* maximal spanning tree. See Figure 3-1a for an example of a nontree graph with a universally embedded path.

Corollary 5. *Let $u, v \in \mathcal{V}$ and $C \subseteq \mathcal{V} \setminus \{u, v\}$ such that $|\mathcal{P}(u; v|C)| > 1$. Suppose*

that $\exists D \subseteq \mathcal{V} \setminus C \setminus \{u, v\}$ *such that* $|\mathcal{P}(u; v|C \cup D)| = 1$. *Then due to* (2.14), *whereby* $I(u; v|C) \geq I(u; v|C \cup D)$, *Proposition 3 implies that universally embedded paths can be used to efficiently compute lower bounds for nonlocal MI between multiply-connected pairs of vertices.*

Definition 6 (Thin Edges). An edge $\{i, j\} \in \mathcal{E}$ of GMRF $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, to which the precision matrix J is Markov, is thin if the corresponding submatrix J_{ij} has exactly one nonzero scalar component. (See Figure 3-1b.)

For vectoral problems, each node may contain a subnetwork of arbitrarily connected scalar random variables (see Figure 3-1b). Under the assumption of thin edges (Definition 6), a unique path between nodes u and v must enter interstitial nodes through one scalar r.v. and leave through one scalar r.v. Therefore, let $\zeta_i(u, v|C) \in (-\infty, 0]$ denote the warped MI between the enter and exit r.v.s of interstitial vectoral node i on $\bar{\pi}_{u:v|C}$, with conditioning set $C \subseteq \mathcal{V} \setminus \{u, v\}$.[1] Note that $\zeta_i(u, v|C)$ can be computed online in $\mathcal{O}(d^3)$ via local marginalization given $\widehat{J}_{\{i\}|C}$, which is an output of GaBP.

Proposition 7 (Thin Vectoral Nonlocal MI Decomposition). *For any GMRF* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *where* $\mathcal{V}$ *indexes random vectors of dimension at most* d *and the edges in* $\mathcal{E}$ *are thin, if* $|\mathcal{P}(u, v|C)| = 1$ *for distinct* $u, v \in \mathcal{V}$ *and* $C \subset \mathcal{V} \setminus \{u, v\}$, *then* $I(\mathsf{x}_u; \mathsf{x}_v|\mathbf{x}_C)$ *can be decomposed as*

$$W(u; v|C) = \sum_{\{i,j\} \in \mathcal{E}(\bar{\pi}_{u:v|C})} W(i; j|C) + \sum_{i \in \bar{\pi}_{u:v|C} \setminus \{u,v\}} \zeta_i(u, v|C). \tag{3.2}$$

3.3 Vectoral Propagation

In Section 3.2, the scalar metric of mutual information was shown to decompose on vectoral networks with "thin" edges, i.e., those that could be parameterized with a single nonzero scalar. For vectoral networks with "thick" edges, mutual information is a scalarization of a fundamentally vectoral relationship between nodes. Therefore,

[1] As node i may have additional neighbors that are not on the u-v path, using the notation $\zeta_i(u, v|C)$ is a convenient way to implicitly specify the enter/exit scalar r.v.s associated with the path. Any unique path subsuming u-v, or any unique path subsumed in u-v for which i is interstitial, will have equivalent ζ_i terms.

Algorithm 3 GAUSSMIPROP$(\mathcal{G}, J, C, r, i, j, Q)$

Require: tree-shaped MRF $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, SPD precision matrix J Markov to $\mathcal{G}$, conditioning set $C \subset \mathcal{V}$, source vertex $r \in \mathcal{V}$, vertices $i, j \in \mathcal{V}$ s.t. $\{i, j\} \in \mathcal{E}$, input matrix $Q \equiv \widehat{J}_{\{r,i\}|C}$

1: $\left[\begin{array}{c|c} D & Y \\ \hline Y^T & Z \end{array}\right] \leftarrow Q$, where $D, Y, Z \in \mathbb{R}^{d \times d}$

2: $\widetilde{D} \leftarrow D - Y(Z - J_{j \to i})^{-1} Y^T$

3: $\widetilde{Y} \leftarrow Y(Z - J_{j \to i})^{-1} J_{ij}$

4: $\widetilde{Z} \leftarrow \widehat{J}_{j \setminus i} - J_{ij}^T (Z - J_{j \to i})^{-1} J_{ij}$

5: $\widetilde{Q} \leftarrow \left[\begin{array}{c|c} \widetilde{D} & \widetilde{Y} \\ \hline \widetilde{Y}^T & \widetilde{Z} \end{array}\right] \quad (\equiv \widehat{J}_{\{r,j\}|C})$

6: $I(r; j|C) \leftarrow -\frac{1}{2} \log\det(\widehat{J}_{\{r\}|C}) - \frac{1}{2} \log\det(\widehat{J}_{\{j\}|C}) + \frac{1}{2} \log\det(\widetilde{Q})$

7: **for** $k \in \Gamma(j|C) \setminus \{i\}$ **do**

8: GAUSSMIPROP$(G, J, C, r, j, k, \widetilde{Q})$

9: **end for**

Algorithm 4 INFOSOURCEQUANT$(\mathcal{G}, J, C, r)$

Require: tree-shaped MRF $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, SPD precision matrix J Markov to $\mathcal{G}$, conditioning set $C \subset \mathcal{V}$, source vertex $r \in \mathcal{V}$

1: Run GaBP with conditioning set C (optional: root r)

2: **for** $i \in \Gamma(r|C)$ **do**

3: $Q \leftarrow \widehat{J}_{\{r,i\}|C}$

4: $I(r; j|C) \leftarrow -\frac{1}{2} \log\det(\widehat{J}_{\{r\}|C}) - \frac{1}{2} \log\det(\widehat{J}_{\{i\}|C}) + \frac{1}{2} \log\det(Q)$

5: **for** $j \in \Gamma(i|C) \setminus \{r\}$ **do**

6: GAUSSMIPROP(G, J, C, r, i, j, Q)

7: **end for**

8: **end for**

one would not expect a similar decomposition of nonlocal MI into the sum of local MI terms. However, the structure of the graph can still be exploited to yield algorithms for efficient information quantification. In this section, a focused information propagation algorithm with the familiar $\mathcal{O}(Nd^3)$ complexity is derived and is shown to generalize the results of the previous section.

Proposition 8 (Vectoral Information Propagation). *Given a tree-shaped MRF $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to which precision matrix J is Markov, for any node $r \in \mathcal{V}$ and conditioning set $C \subset \mathcal{V} \setminus \{r\}$, Algorithm 4 computes the set of nonlocal pairwise mutual information terms $\{I(r; i|C)\}_{i \in \mathcal{V} \setminus C}$ with overall complexity $\mathcal{O}(Nd^3)$.*

The decompositions of Section 3.2 are special cases of Proposition 8.

Corollary 9. *For any unique vectoral path $\bar{\pi}_{u:v|C}$ with interstitial node $k \in \mathcal{V}$, then $W(u;v|C) = W(u;k|C) + W(k;v|C) + \zeta_k(u;v|C)$ if k contains a single scalar component whose removal completely separates u and v.*

3.4 Efficient Focused Greedy Selection

The nonlocal MI decompositions of Section 3.2 can be used to efficiently solve the focused greedy selection problem, which at each iteration, given the subset $\mathcal{A} \subset \mathcal{S}$ of previously selected observable random variables, is

$$\underset{\{y \in \mathcal{S}\setminus\mathcal{A} \,:\, c(y) \leq \beta - c(\mathcal{A})\}}{\operatorname{argmax}} I(\mathbf{x}_{\mathcal{R}}; \mathbf{x}_y \mid \mathbf{x}_{\mathcal{A}}).$$

To proceed, first consider the singleton case $\mathcal{R} = \{r\}$ for $r \in \mathcal{U}$. The results of Proposition 8 indicate that all scores $\{I(r;y|\mathcal{A})\}_{y \in \mathcal{S}\setminus\mathcal{A}}$ can be can collectively be computed at each iteration of the greedy algorithm with overall complexity $\mathcal{O}(Nd^3)$.

Now consider $|\mathcal{R}| > 1$. Let $R = (r_1, \ldots, r_{|\mathcal{R}|})$ be an ordering of the elements of $\mathcal{R}$, and let R_k be the first k elements of R. Then, by the chain rule of mutual information, $I(\mathcal{R}; y \mid \mathcal{A}) = \sum_{k=1}^{|\mathcal{R}|} I(r_k; y \mid \mathcal{A} \cup R_{k-1})$, $y \in \mathcal{S} \setminus \mathcal{A}$, where each term in the sum is a pairwise (potentially nonlocal) MI evaluation. The implication is that one can run $|\mathcal{R}|$ separate instances of GaBP, each using a different conditioning set $\mathcal{A} \cup R_{k-1}$, to compute the initial message set used by Algorithm 4. The chain rule suggests one should then sum the MI scores of these $|\mathcal{R}|$ instances to yield the scores $I(\mathcal{R}; y|\mathcal{A})$ for $y \in \mathcal{S} \setminus \mathcal{A}$. The total cost of a greedy update is then $\mathcal{O}\left(N|\mathcal{R}|d^3\right)$.

3.4.1 Distributed Implementation

One of the benefits of the focused greedy selection algorithm is its amenability to parallelization. All quantities needed to form the initial message set are derived from GaBP, which is parallelizable and guaranteed to converge on trees in at most $\operatorname{diam}(\mathcal{G})$ iterations [91]. Parallelization reallocates the expense of quantification across net-

worked computational resources, often leading to faster solution times and enabling larger problem instantiations than are otherwise permissible. However, full parallelization, wherein each node $i \in \mathcal{V}$ is viewed as separate computing resource, incurs a multiplicative overhead of $\mathcal{O}(\text{diam}(\mathcal{G}))$ due to each i having to send $|\Gamma(i)|$ messages $\text{diam}(\mathcal{G})$ times, yielding local communication costs of $\mathcal{O}(\text{diam}(\mathcal{G})|\Gamma(i)| \cdot d^3)$ and overall complexity of $\mathcal{O}(\text{diam}(\mathcal{G}) \cdot N|\mathcal{R}|d^3)$. This overhead can be fractionally alleviated by instead assigning to every computational resource a connected subgraph of $\mathcal{G}$ of appropriate cardinality.

3.4.2 Serial Implementation

It should also be noted that if the quantification is instead performed using serial BP — which can be conceptualized as choosing an arbitrary root, collecting messages from the leaves up to the root, and disseminating messages back down again — a factor of 2 savings can be achieved for $R_2, \ldots, R_{|\mathcal{R}|}$ by noting that in moving between instances k and $k+1$, only r_k is added to the conditioning set [47]. Therefore, by reassigning r_k as the root for the BP instance associated with r_{k+1} (i.e., $\mathcal{A} \cup R_k$ as the conditioning set), only the second half of the message passing schedule (disseminating messages from the root to the leaves) is necessary. This computational trick that reuses previous BP results is subsequently referred to as "caching."

3.5 Experiments

To benchmark the runtime performance of the algorithm in Section 3.4, its serial GaBP variant was implemented in Java, with and without the caching trick described above. The algorithm of Section 3.4 is compared here with greedy selectors that use matrix inversion (with cubic complexity) to compute nonlocal mutual information measures. The inversion-based quantifiers, which are described in the two subsections that follow, were implemented using Colt sparse matrix libraries [13].

3.5.1 Naïve Inversion

Whenever a mutual information term of the form $I(A;B|C)$ is needed, the procedure `NaïveInversion` conditions J on C and computes the marginal covariance matrices $P_{A\cup B|C}$, $P_{A|C}$, and $P_{B|C}$ of (2.25) using standard matrix inversion, which is $\mathcal{O}(N^3d^3)$. A greedy selection update, which requires computing marginal information gain scores $\{I(\mathcal{R};y|\mathcal{A})\}_{y\in\mathcal{S}\setminus\mathcal{A}}$, thereby requires $\mathcal{O}(N^3|\mathcal{S}|d^3)$ operations using this procedure.

3.5.2 Block Inversion

Intuitively, the `NaïveInversion` procedure appears wasteful even for an inversion-based method, as it repeats many of the marginalization operations needed to form $\{I(\mathcal{R};y|\mathcal{A})\}_{y\in\mathcal{S}\setminus\mathcal{A}}$. The `BlockInversion` procedure attempts to rectify this. Given a previous selection set $\mathcal{A}$, `BlockInversion` conditions J on $\mathcal{A}$ and marginalizes out nuisances $\mathcal{U}\setminus\mathcal{R}$ (along with infeasible observation selections $\{y\in\mathcal{S}\setminus\mathcal{A} \mid c(y) > \beta - c(\mathcal{A})\}$) using Schur complements. The complexity of this approach, for each greedy update, is $\mathcal{O}((|\mathcal{S}|^4 + |\mathcal{R}||\mathcal{S}|^3 + |\mathcal{R}|^3|\mathcal{S}| + N^3)d^3)$. `BlockInversion` has the same worst-case asymptotic complexity of $\mathcal{O}(N^3|\mathcal{S}|d^3)$ as `NaïveInversion` but may achieve a significant reduction in computation depending on how $|\mathcal{R}|$ and $|\mathcal{S}|$ scale with N.

3.5.3 Runtime Comparison

Figure 3-2 shows the comparative mean runtime performance of each of the quantifiers for scalar networks of size N, where the mean is taken over the 20 problem instances proposed for each value of N. Each problem instance consists of a randomly generated, symmetric, positive-definite, tree-shaped precision matrix J, along with a randomly labeled $\mathcal{S}$ (such that, arbitrarily, $|\mathcal{S}| = 0.3|\mathcal{V}|$) and $\mathcal{R}$ (such that $|\mathcal{R}| = 5$), as well as randomly selected budget and heterogeneous costs defined over $\mathcal{S}$. Note that all selectors return the same greedy selection; this comparison concerns how the decompositions proposed in this thesis aid in the computational performance. In the figure, it is clear that the GaBP-based quantification algorithms of Section 3.4 vastly

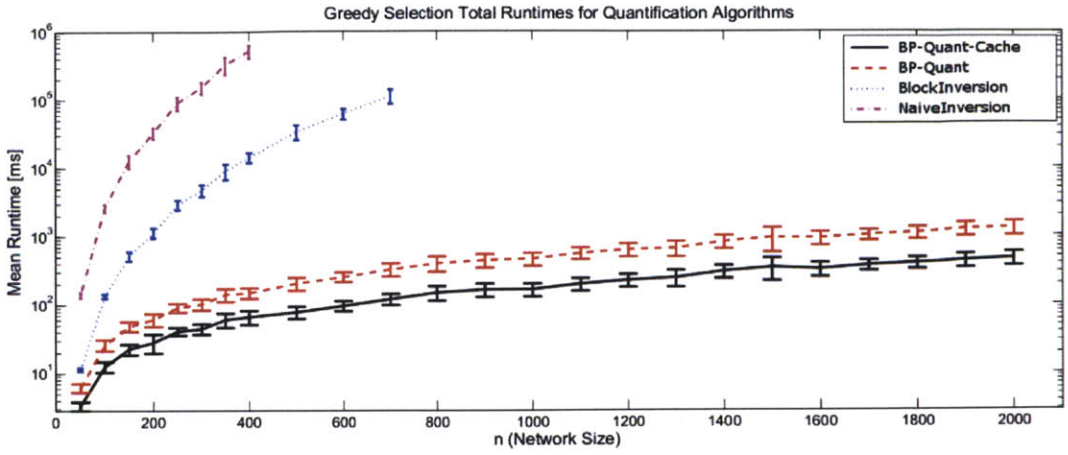


Figure 3-2: Performance of GaBP-based and inversion-based quantifiers used in greedy selectors. For each N, the mean of the runtimes over 20 random scalar problem instances is displayed. The `BP-Quant` algorithm of Section 3.4 empirically has approximately linear complexity; caching reduces the mean runtime by a factor of approximately 2.

outperform both inversion-based methods; for relatively small N, the solution times for the inversion-based methods became prohibitively long. Conversely, the behavior of the BP-based quantifiers empirically confirms the asymptotic $\mathcal{O}(N)$ complexity of this chapter's method for scalar networks.

3.6 Proofs

In order to prove Proposition 3, it is convenient to first prove the following lemma.

Lemma 10. *Consider a connected GMRF $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ parameterized by precision matrix J Markov to $\mathcal{G}$ and a unique path $\bar{\pi}$ embedded in $\mathcal{G}$. The marginal precision matrix $\widehat{J}_{\bar{\pi}}$ has block off-diagonal elements identical to those in the submatrix of J corresponding to variables in $\bar{\pi}$, and block diagonal elements that are the Schur complements of the submatrices corresponding to the sidegraphs separated by $\bar{\pi}$.*

Proof. Assume without loss of generality that the unique path under consideration is $(1, 2, \ldots, k-1, k)$. Because $\bar{\pi}_{1:k}$ is unique, the graph $\tilde{\mathcal{G}}$ induced by $\mathcal{V} \setminus \bar{\pi}_{1:k}$ can be thought of as the union of conditionally independent "sidegraphs" $\tilde{\mathcal{G}}_1, \ldots, \tilde{\mathcal{G}}_k$, each of

which is connected in $\mathcal{G}$ to a single node in $\bar{\pi}_{1:k}$ (see Figure 3-1a). Let $J_{\bar{\pi}_{1:k}}$ denote the (block tridiagonal) matrix parameterizing the joint potential (i.e., the product of singleton and edge potentials in the factorization of the full joint distribution of $\mathbf{x}$) over the chain $(1,\ldots,k)$. For all $i \in \{1,\ldots,k\}$, let $J_{i,\tilde{\mathcal{G}}_i}$ be the matrix parameterizing the potentials over edges between i and $\Gamma(i)\setminus\bar{\pi}_{1:k}$. Likewise, let $J_{\tilde{\mathcal{G}}_i}$ denote the matrix parameterizing the joint potential over the subgraph $\tilde{\mathcal{G}}_i$.

Now consider a permutation to the last $N-k$ components of J such that $J_{\tilde{\mathcal{G}}_1},\ldots,J_{\tilde{\mathcal{G}}_k}$ are ordered as such, whereby

$$J \triangleq \left[\begin{array}{c|c} J_{\bar{\pi}_{1:k}} & J_{\bar{\pi}_{1:k},\tilde{\mathcal{G}}} \\ \hline J^T_{\bar{\pi}_{1:k},\tilde{\mathcal{G}}} & J_{\tilde{\mathcal{G}}} \end{array}\right].$$

In this permuted matrix, $J_{\tilde{\mathcal{G}}}$ is block diagonal – due to conditional independence of the sidegraphs – with elements $J_{\tilde{\mathcal{G}}_i}$. Similarly, the upper-right block submatrix $J_{\bar{\pi}_{1:k},\tilde{\mathcal{G}}}$ is also block diagonal with elements $J_{i,\tilde{\mathcal{G}}_i}$. Thus, the marginal distribution $p_{\mathbf{x}_1,\ldots,\mathbf{x}_k}$ is parameterized by a precision matrix

$$\widehat{J}_{\bar{\pi}_{1:k}} = J_{\bar{\pi}_{1:k}} - J_{\bar{\pi}_{1:k},\tilde{\mathcal{G}}} J^{-1}_{\tilde{\mathcal{G}}} J^T_{\bar{\pi}_{1:k},\tilde{\mathcal{G}}},$$

where the subtractive term is a product of block diagonal matrices and, thus, is itself a block diagonal matrix. Therefore, the marginal precision matrix $\widehat{J}_{\bar{\pi}_{1:k}}$ has block off-diagonal elements identical to those of the submatrix $J_{\bar{\pi}_{1:k}}$ of the (full) joint precision matrix; each block diagonal element is the Schur complement of each $J_{\tilde{\mathcal{G}}_i}$, $i = 1,\ldots,k$. □

Remark 11. Lemma 10 implies that if Proposition 3 holds for any chain of length k between nodes u and v, it must also hold for the more general class of graphs in which $|\mathcal{P}(u,v)| = 1$ (i.e., there is a unique path between u and v, but there are sidegraphs attached to each vertex in the path). Therefore, one need only prove Proposition 3 for chains of arbitrary length. Furthermore, conditioning only severs nodes from the graph component considered; provided C is not a separator for u, v, in which case $I(\mathsf{x}_u;\mathsf{x}_v|\mathbf{x}_C) = 0$, one need only prove Proposition 3 for the case where $C = \emptyset$.

Proof of Proposition 3. The proof proceeds by induction on the length k of the chain. The base case considered is a chain of length $k = 3$, for which the precision matrix is

$$J = \begin{bmatrix} J_{11} & J_{12} & 0 \\ J_{12} & J_{22} & J_{23} \\ 0 & J_{23} & J_{33} \end{bmatrix}.$$

By Remark 2, one need only show that $|\rho_{13}| = |\rho_{12}| \cdot |\rho_{23}|$. The covariance matrix

$$J^{-1} = \frac{1}{\det(J)} \begin{bmatrix} J_{22}J_{33} - J_{23}^2 & -J_{12}J_{33} & J_{12}J_{23} \\ -J_{12}J_{33} & J_{11}J_{33} & -J_{11}J_{23} \\ J_{12}J_{23} & -J_{11}J_{23} & J_{11}J_{22} - J_{12}^2 \end{bmatrix}$$

can be used to form the correlation coefficients $\rho_{ij} = P_{ij}/\sqrt{P_{ii}P_{jj}}$, where one can confirm that

$$\begin{aligned} \rho_{12} &= -J_{12}J_{33}/\sqrt{J_{11}J_{33}(J_{22}J_{33} - J_{23}^2)} \\ \rho_{23} &= -J_{11}J_{23}/\sqrt{J_{11}J_{33}(J_{11}J_{22} - J_{12}^2)} \\ \rho_{13} &= J_{12}J_{23}/\sqrt{(J_{11}J_{22} - J_{12}^2)(J_{22}J_{33} - J_{23}^2)} \\ &= \rho_{12} \cdot \rho_{23}, \end{aligned}$$

thus proving the base case.

Now, assume the result of the proposition holds for a unique path $\bar{\pi}_{1:k}$ of length k embedded in graph $\mathcal{G}$, and consider node $k+1 \in \Gamma(k) \setminus \bar{\pi}_{1:k}$. By Lemma 10, one can restrict attention to the marginal chain over $(1, \ldots, k, k+1)$. Pairwise decomposition on subchains of length k yields an expression for $I(\mathsf{x}_1; \mathsf{x}_k)$, which by (4) can alternatively be expressed in terms of the determinants of marginal precision matrices. Therefore, if one marginalizes nodes in $\{2, \ldots, k-1\}$ (under any elimination ordering), one is left with a graph over nodes 1, k, and $k+1$. The MI between k and $k+1$, which are adjacent in $\mathcal{G}$, can be computed from the local GaBP messages comprising the marginal node and edge precision matrices. On this 3-node network,

for which the base case holds,

$$\begin{aligned} W(1;k+1) &= W(k;k+1) + W(1;k) \\ &= W(k;k+1) + \sum_{\{i,j\}\in\mathcal{E}(\bar{\pi}_1)k} W(i;j) \\ &= \sum_{\{i,j\}\in\mathcal{E}(\bar{\pi}_1)k+1} W(i;j). \end{aligned}$$

Therefore, Proposition 3 holds for both chains and, by Lemma 10, unique paths of arbitrary length in GMRFs with scalar variables. □

Proof of Proposition 7. The proof follows closely that of Proposition 3. By Lemma 10, assume without loss of generality that the unique path under consideration is a vectoral chain of length k with sequentially indexed nodes, i.e., $(1, 2, \ldots, k-1, k)$. Thinness of edges $\{i, j\} \in \mathcal{E}$ implies $W(i;j) = \log|\rho_{ij}|$, as before. Let $i \in \{2, \ldots, k-1\}$ be an arbitrary interstitial node. On section $(i-1, i, i+1)$ of the chain, thinness of $\{i-1, i\}$ and $\{i, i+1\}$ implies that on $\bar{\pi}_{1:k}$, there exists one inlet to and one outlet from i. Let m and q denote the column of $J_{i-1,i}$ and row of $J_{i,i+1}$, respectively, containing nonzero elements. Then the path through the internal structure of node i can be simplified by marginalizing out, as computed via Schur complement from $\widehat{J}_{\{i\}}$ in $\mathcal{O}(d^3)$, all scalar elements of i except m and q. Thus, $\zeta_i(u, v)$ is merely the warped mutual information between m and q, and problem reduces to a scalar chain with alternating W and ζ terms. □

Proof of Proposition 8. As in the proof of Proposition 3, consider a universally embedded chain of length m and assume w.l.o.g. that the nodes in the chain are labeled lexicographically $(1, 2, \ldots, m)$. The marginal distribution over $(\mathsf{x}_1, \mathsf{x}_2, \mathsf{x}_3)$ is parameterized by the precision matrix

$$\widehat{J}_{\{1,2,3\}} = \begin{bmatrix} \widehat{J}_{1\backslash 2} & J_{12} & \mathbf{0} \\ J_{12}^T & \widehat{J}_{2\backslash\{1,3\}} & J_{23} \\ \mathbf{0} & J_{23}^T & \widehat{J}_{3\backslash 2} \end{bmatrix} \in \mathbb{R}^{3d\times 3d}, \tag{3.3}$$

where $J_{ij} = (J)_{ij}$ and $\widehat{J}_{i\backslash A} = J_{ii} + \sum_{k\in\Gamma(i)\backslash A} J_{k\to i}$. The (nonlocal) marginal distribution over $(\mathsf{x}_1, \mathsf{x}_3)$ can then be computed in $\mathcal{O}(d^3)$ via Schur complement, i.e,

$$\widehat{J}_{\{1,3\}} = \begin{bmatrix} \widehat{J}_{1\backslash 2} & 0 \\ 0 & \widehat{J}_{3\backslash 2} \end{bmatrix} - \begin{bmatrix} J_{12} \\ J_{23}^T \end{bmatrix} \left(\widehat{J}_{2\backslash\{1,3\}}\right)^{-1} \begin{bmatrix} J_{12}^T & J_{23} \end{bmatrix}. \tag{3.4}$$

Considering the marginal chain over $(1, 3, 4, \ldots, m)$ parameterized by $\widetilde{J}$. The marginal distribution of $(\mathsf{x}_1, \mathsf{x}_3, \mathsf{x}_4)$ is parameterized by

$$\widehat{J}_{\{1,2,3\}} = \begin{bmatrix} \widehat{J}_{1\backslash 3} & \widetilde{J}_{13} & 0 \\ \widetilde{J}_{13}^T & \widehat{J}_{3\backslash\{1,4\}} & J_{34} \\ 0 & J_{34}^T & \widehat{J}_{4\backslash 3} \end{bmatrix} \in \mathbb{R}^{3d\times 3d}. \tag{3.5}$$

However,

$$Q \triangleq \widehat{J}_{\{1,3\}} = \begin{bmatrix} \widehat{J}_{1\backslash 3} & \widetilde{J}_{13} \\ \widetilde{J}_{13}^T & \widehat{J}_{3\backslash 1} \end{bmatrix}. \tag{3.6}$$

By repeated application of the marginalization, the update rules of Algorithm 3 follow immediately. Note that each call to Algorithm 3 entails $\mathcal{O}(1)$ updates of complexity $\mathcal{O}(d^3)$, and the recursion on a tree terminates after at most $N-1$ edges have been traversed. Thus, the overall expense of Algorithm 4 is $\mathcal{O}(Nd^3)$, after which one has access to all pairwise MI terms $\{I(r;i|C)\}_{i\in\mathcal{V}\backslash C}$ between source $r \in \mathcal{V}$ and each other vertex.

□

Chapter 4

Information Quantification in Loopy Gaussian Graphical Models

For Gaussian belief networks with large N (many vertices), evaluating the MI objective in (1.1) via matrix inversion is cubic in N, which may be prohibitively expensive. Chapter 3 presented an efficient, *exact* algorithm for reducing the complexity of pairwise nonlocal MI evaluations on Gaussian trees to $\mathcal{O}(Nd^3)$, i.e., linear in the number of vertices. The main objective of this chapter is providing a similar reduction in complexity for loopy Gaussian graphical models. The presented approach is an *iterative* approximation algorithm for computing (to within a specified tolerance) specific components of conditional covariance matrices, from which MI can be computed. The complexity per iteration is linear in N, and the convergence will be demonstrated (in Section 4.3) to often be subquadratic in N, leading to a relative asymptotic efficiency over naïve linear algebraic techniques.

4.1 Background

4.1.1 Embedded Trees

The embedded trees (ET) algorithm was introduced in [83, 90] to iteratively compute both conditional means and marginal error variances in Gaussian graphical models

with cycles. Although the algorithm requires only the identification of subgraphs on which inference is tractable, and extensions to, for example, embedded polygons [21] and embedded hypergraphs [14] have been considered, we will focus for clarity of discussion on embedded trees.

Let $\mathbf{x} \sim \mathcal{N}^{-1}(\mathbf{0}, J)$ be a Gaussian distributed random vector Markov to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that contains cycles. Consider an alternatively distributed random vector $\mathbf{x}_\mathcal{T} \sim \mathcal{N}^{-1}(\mathbf{0}, J_\mathcal{T})$ that is of the same dimension as $\mathbf{x}$ but is instead Markov to a cycle-free subgraph $\mathcal{G}_\mathcal{T} = (\mathcal{V}, \mathcal{E}_\mathcal{T})$ of $\mathcal{G}$ (in the sense that $\mathcal{E}_\mathcal{T} \subset \mathcal{E}$). The tree-shaped (and symmetric, positive definite) inverse covariance matrix $J_\mathcal{T}$ can be decomposed as $J_\mathcal{T} = J + K_\mathcal{T}$, where $K_\mathcal{T}$ is any symmetric *cutting matrix* that enforces the sparsity pattern of $J_\mathcal{T}$ by zeroing off-diagonal elements of J corresponding to cut edges $\mathcal{E} \setminus \mathcal{E}_\mathcal{T}$. Since many cutting matrices $K_\mathcal{T}$ will result in a tree-shaped inverse covariance $J_\mathcal{T}$ Markov to $\mathcal{G}_\mathcal{T}$, we will restrict attention to so-called *regular* cutting matrices, whose nonzero elements are constrained to lie at the intersection of the rows and columns corresponding to the vertices incident to cut edges. Note that $K_\mathcal{T}$ can always be chosen such that $\text{rank}(K_\mathcal{T})$ is at most $\mathcal{O}(Ed)$, where we will use $E \triangleq |\mathcal{E} \setminus \mathcal{E}_\mathcal{T}|$ to denote the number of cut edges.

Conditional Means

Given an initial solution $\hat{\mathbf{x}}^{(0)}$ to (2.18), the single-tree Richardson iteration [96] induced by embedded tree $\mathcal{G}_\mathcal{T}$ with cutting matrix $K_\mathcal{T}$ and associated inverse covariance $J_\mathcal{T} = J + K_\mathcal{T}$ is

$$\hat{\mathbf{x}}^{(n)} = J_\mathcal{T}^{-1} \left(K_\mathcal{T} \hat{\mathbf{x}}^{(n-1)} + \mathbf{h} \right). \tag{4.1}$$

Thus, each update $\hat{\mathbf{x}}^{(n)}$ is the solution of a synthetic inference problem (2.18) with precision matrix $\widetilde{J} = J_\mathcal{T}$ and potential vector $\widetilde{\mathbf{h}} = K_\mathcal{T} \hat{\mathbf{x}}^{(n-1)} + \mathbf{h}$. This update requires a total of $\mathcal{O}(Nd^3 + Ed^2)$ operations, where $\mathcal{O}(Nd^3)$ is due to solving $\widetilde{J} \hat{\mathbf{x}}^{(n)} = \widetilde{\mathbf{h}}$ with a tree-shaped graph, and where $\mathcal{O}(Ed^2)$ with $E = |\mathcal{E} \setminus \mathcal{E}_\mathcal{T}|$ is due to forming $\widetilde{\mathbf{h}}$. In the case that E is at most $\mathcal{O}(N)$, the overall complexity *per iteration* is $\mathcal{O}(Nd^3)$.

Letting $\rho(D) \triangleq \max_{\lambda \in \{\lambda_i(D)\}} |\lambda|$ denote the spectral radius of a square matrix D, the asymptotic convergence rate of the single-tree iteration (4.1) is

$$\rho(J_{\mathcal{T}}^{-1} K_{\mathcal{T}}) = \rho(I - J_{\mathcal{T}}^{-1} J), \tag{4.2}$$

with convergence to $\hat{\mathbf{x}}$ guaranteed (regardless of $\hat{\mathbf{x}}^{(0)}$) if and only if $\rho(J_{\mathcal{T}}^{-1} K_{\mathcal{T}}) < 1$. Inherent in Equations (4.1) and (4.2) is a tradeoff in the choice of embedded structure between the tractability of solving $J_{\mathcal{T}} \hat{\mathbf{x}}^{(n)} = \widetilde{\mathbf{h}}$ and the approximation strength of $J_{\mathcal{T}} \approx J$ for fast convergence.

The ET algorithm [85] is conceptualized as a nonstationary Richardson iteration with multiple matrix splittings of J. Let $\{G_{\mathcal{T}_n}\}_{n=1}^{\infty}$ be sequence of embedded trees within $\mathcal{G}$, and let $\{K_{\mathcal{T}_n}\}_{n=1}^{\infty}$ a sequence of cutting matrices such that $J_{\mathcal{T}_n} = J + K_{\mathcal{T}_n}$ is Markov to $\mathcal{G}_{\mathcal{T}_n}$ for $n = 1, \ldots, \infty$. The nonstationary Richardson update is then

$$\hat{\mathbf{x}}^{(n)} = J_{\mathcal{T}_n}^{-1} \left(K_{\mathcal{T}_n} \hat{\mathbf{x}}^{(n-1)} + \mathbf{h} \right), \tag{4.3}$$

with error $e^{(n)} \triangleq \hat{\mathbf{x}}^{(n)} - \hat{\mathbf{x}}$ that evolves according to

$$e^{(n)} = J_{\mathcal{T}_n}^{-1} K_{\mathcal{T}_n} e^{(n-1)}. \tag{4.4}$$

The criterion for convergence is when the normalized residual error $||K_{\mathcal{T}_n}(\hat{\mathbf{x}}^{(n)} - \hat{\mathbf{x}}^{(n-1)})||_2 / ||\mathbf{h}||_2$ falls below a specified tolerance $\epsilon > 0$. The sparsity of $K_{\mathcal{T}_n}$ permits the efficient computation of this residual.

When $\{G_{\mathcal{T}_n}, K_{\mathcal{T}_n}\}_{n=1}^{\infty}$ is periodic in n, a convergence rate analysis similar to (4.2) is given in [85]. It is also demonstrated that using multiple embedded trees can significantly improve the convergence rate. Online adaptive selection of the embedded tree was explored in [14] by scoring edges according to single-edge walk-sums and forming a maximum weight spanning tree in $\mathcal{O}(|\mathcal{E}| \log |N|)$.

Marginal Variances

Given that $\text{rank}(K_{\mathcal{T}}) \leq 2Ed$ [83], where $E = |\mathcal{E} \setminus \mathcal{E}_{\mathcal{T}}|$, an additive rank-one decomposition

$$K_{\mathcal{T}} = \sum_i w_i u_i u_i^T, \qquad u_i \in \mathbb{R}^{Nd} \tag{4.5}$$

can be substituted in the fixed-point equation [85]

$$P = J_{\mathcal{T}}^{-1} + J_{\mathcal{T}}^{-1} K_{\mathcal{T}} P, \tag{4.6}$$

yielding

$$P = J_{\mathcal{T}}^{-1} + \sum_i w_i (J_{\mathcal{T}}^{-1} u_i)(P u_i)^T. \tag{4.7}$$

Solving for the vertex-marginal covariances $P_i = P_{ii}, i \in \mathcal{V}$ which are the block-diagonal entries of P, requires:

- solving for the block-diagonal entries of $J_{\mathcal{T}}^{-1}$, with one-time complexity $\mathcal{O}(Nd^3)$ via GaBP;
- solving the synthetic inference problems $J_{\mathcal{T}} z_i = u_i$, for all $\mathcal{O}(Ed)$ vectors u_i of $K_{\mathcal{T}}$ in (4.5), with one-time total complexity $\mathcal{O}(Nd^3 \cdot Ed) = \mathcal{O}(NEd^4)$ via GaBP;
- solving the synthetic inference problems $Jz_i = u_i$, for all $\mathcal{O}(Ed)$ vectors u_i of $K_{\mathcal{T}}$ in (4.5), with *per iteration* total complexity of $\mathcal{O}(NEd^4)$ operations via ET conditional means (4.3);
- assembling the above components via (4.7).

Note that there exists an alternative decomposition (4.5) into $\mathcal{O}(Wd)$ rank-one matrices using a cardinality-W vertex cover of $\mathcal{E} \setminus \mathcal{E}_{\mathcal{T}}$ (where $W \leq E$ for any minimal vertex cover); this alternative decomposition requires solving a symmetric quadratic eigenvalue problem [85].

4.1.2 Competing Methods

Other methodologies have been proposed to perform inference in loopy graphs. Loopy belief propagation is simply parallel belief propagation performed on graphs with cycles; if it converges, it does so to the correct mean but, in general, to incorrect variances [91]. Extended message passing augments the original BP messages and provides for convergence to the correct variances, but its complexity is $\mathcal{O}(NL^2)$ in the scalar case, where L is the number of vertices incident to any cut edge, and it requires the full message schedule to be executed to produce an estimate [69]. Linear response algorithms can be used to compute pairwise marginal distributions for nonadjacent pairs of vertices, but at a complexity of $\mathcal{O}(N|\mathcal{E}|d^3)$, which may be excessive for large N and $|\mathcal{E}| = \mathcal{O}(N)$ given that conditional MI requires only very specific pairwise marginals [92].

It is also possible to perform efficient inference if it is known that the removal of a particular subset of $\mathcal{V}$, called a feedback vertex set (FVS), will induce a tree-shaped subgraph [53]. The resulting belief propagation-like inference algorithm, called feedback message passing (FMP), has complexity $\mathcal{O}(Nk^2)$ for scalar networks, where k is the cardinality of the FVS. If the topological structure of the graphical model is well known *a priori*, or if the graph is learned by an algorithm oriented towards forming FVSs [54], then identification of an FVS is straightforward. Otherwise, it may be computationally expensive to subsequently identify an FVS of reasonable size; in contrast, finding a spanning tree (as the ET algorithm does) is comparatively simple. Moreover, for large k, standard FMP might be prohibitively expensive. In such cases, one may form a subset of vertices called a pseudo-FVS, which does not necessarily induce a tree-shaped subgraph upon its removal, but for which approximate FMP returns the exact marginal variances among the pseudo-FVS and inaccurate variances elsewhere [54]. While FMP relies on a distributed message passing algorithm for the tree-shaped subgraph and a centralized algorithm among the feedback vertices, Recursive FMP [55] provides protocols to fully distribute both message passing subroutines.

4.2 ET Mutual Information Quantification (ET-MIQ)

This section describes the application of embedded trees to efficient iterative computation of nonlocal mutual information measures on loopy Gaussian graphs.

It is typically intractable to enumerate all possible selection sets $\mathcal{A} \in 2^{\mathcal{V}}$ and evaluate the resulting MI objective $I(\mathcal{R};\mathcal{A})$. Often, one balances tractability with performance by using suboptimal selection heuristics with either a priori or online-computable performance bounds [39, 47]. Starting from an empty selection $\mathcal{A} \leftarrow \emptyset$, the greedy heuristic (cf. Section 2.5)

$$\begin{aligned} a &\leftarrow \underset{\{y \in \mathcal{S}\setminus\mathcal{A} \,:\, c(y) \le \beta - c(\mathcal{A})\}}{\operatorname{argmax}} I(\mathcal{R}; y|\mathcal{A}) \\ \mathcal{A} &\leftarrow \mathcal{A} \cup \{a\} \end{aligned} \tag{4.8}$$

selects one unselected observable variable with the highest marginal increase in objective and continues to do so until the budget is expended. By comparison to Equation (2.25), the MI evaluations needed to perform a greedy update are of the form

$$I(\mathcal{R}; y|\mathcal{A}) = \frac{1}{2} \log \frac{\det(P_{\mathcal{R}|\mathcal{A}}) \det(P_{\{y\}|\mathcal{A}})}{\det(P_{\mathcal{R}\cup\{y\}|\mathcal{A}})} \tag{4.9}$$

While inverse covariance matrices obey specific sparsity patterns, covariance matrices are generally dense. Thus two of the determinants in Equation (4.9) require $\mathcal{O}(|\mathcal{R}|^3 d^3)$ operations to compute. If $|\mathcal{R}|$ is $\mathcal{O}(N)$ (e.g., the graph represents a regular pattern, a constant fraction of which is to be inferred), then such determinants would be intractable for large N. We instead fix some ordering R over the elements of $\mathcal{R}$, denoting by r_k its kth element, $R_k = \cup_{i=1}^{k} \{r_i\}$ its first k elements, and appeal to the chain rule of mutual information:

$$\begin{aligned} I(\mathcal{R}; y \mid \mathcal{A}) = I(r_1; y \mid \mathcal{A}) + I(r_2; y \mid \mathcal{A} \cup R_1) + \ldots \\ + I(r_{|\mathcal{R}|}; y \mid \mathcal{A} \cup R_{|\mathcal{R}|-1}). \end{aligned} \tag{4.10}$$

The advantage of this expansion is twofold. Each term in the summation is a pairwise mutual information term. Given an efficient method for computing marginal covariance matrices (the focus of the remainder of this section), the determinants in (2.25) can be evaluated in $\mathcal{O}(d^3)$ operations. More pressingly, conditioning in an undirected graphical model removes paths from the graph (by the global Markov property), potentially simplifying the structure over which one must perform the quantification. Therefore, the chain rule converts the problem of evaluating a set mutual information measure $I(\mathcal{R}; y|\mathcal{A})$ into $|\mathcal{R}|$ separate pairwise MI computations that *decrease* in difficulty as the conditioning set expands.

It suffices to describe how to compute *one* of the $|\mathcal{R}|$ terms in the summation (4.10); the template will be repeated for the other $|\mathcal{R}| - 1$ terms, but with a modified conditioning set. In the remainder of this section, we show how to efficiently compute $I(r; y|C)$ for all $y \in \mathcal{S} \setminus C$ provided some $r \in \mathcal{R}$ and conditioning set $C \subset \mathcal{V} \setminus \{r\}$. Since conditioning on C can be performed by selecting the appropriate submatrix of J corresponding to $\mathcal{V} \setminus C$, we will assume for clarity of presentation and without loss of generality[1] that either $C = \emptyset$ or that we are always working with a J resulting from a larger J' that has been conditioned on C. The resulting MI terms, in further simplification of (4.9), are of the form

$$I(r; y) = \frac{1}{2} \log \frac{\det(P_{\{r\}}) \det(P_{\{y\}})}{\det(P_{\{r,y\}})}, \tag{4.11}$$

where $P_{\{r\}} = P_{rr}$ and $P_{\{y\}} = P_{yy}$ are the $d \times d$ marginal covariances on the diagonal, and where $P_{\{r,y\}}$ is the $2d \times 2d$ block submatrix of the (symmetric) covariance $P = J^{-1}$:

$$P_{\{r,y\}} = \begin{bmatrix} P_{rr} & P_{ry} \\ (P_{ry})^T & P_{yy} \end{bmatrix}.$$

In addition to the marginal covariances on the diagonal, the $d \times d$ off-diagonal cross-covariance term $P_{ry} = (P_{yr})^T$ is needed to complete $P_{\{r,y\}}$. If it were possible to

[1] Alternatively, the unconditioned J can be used by treating conditioned vertices as blocked (not passing messages) and by zeroing the elements of $\mathbf{h}$ and $\hat{\mathbf{x}}^{(n)}$ in (4.1) corresponding to C.

efficiently estimate the d columns of P corresponding to r, all such cross-covariance terms $P_{ry}, \forall y \in \mathcal{S}$, would be available. Therefore, let P be partitioned into columns $\{p_i\}_{i=1}^{Nd}$ and assume without loss of generality that r corresponds to $p_1, \ldots, p_d$. Let e_i be the ith Nd-dimensional axis vector (with a 1 in the ith position). Then $p_i \equiv Pe_i, i = 1, \ldots, d$, can be estimated using the synthetic inference problem

$$Jp_i = e_i. \tag{4.12}$$

Thus, by comparison to (2.18) and (4.3), the first d columns of P can be estimated with a complexity of $\mathcal{O}(Nd^4)$ per ET iteration.

Using the results of Section 4.1.1, the marginal variances can be estimated in $\mathcal{O}(NEd^4)$ per iteration, where $E = |\mathcal{E} \setminus \mathcal{E}_{\mathcal{T}_n}|$ is the number of cut edges. One can subsequently form each matrix $P_{\{r,y\}}$, $y \in \mathcal{S}$ in $\mathcal{O}(d^2)$ and take its determinant in $\mathcal{O}(d^3)$. Since $|\mathcal{S}| < N$, the ET-MIQ procedure outlined in the section can be used to iteratively estimate the set $\{I(r;y)\}_{y\in\mathcal{S}}$ with total complexity $\mathcal{O}(NEd^4)$ operations per iteration. Returning to the greedy selection of Equation (4.8) and the chain rule of (2.12), given a subset $\mathcal{A} \subset \mathcal{S}$ of previous selections, the set of marginal gains $\{I(\mathcal{R};y|\mathcal{A})\}_{y\in\mathcal{S}\setminus\mathcal{A}}$ can be estimated in $\mathcal{O}(N|\mathcal{R}|Ed^4)$ operations per iteration.

4.3 Experiments

4.3.1 Alternative Methods

In order to demonstrate the comparative performance of the ET-MIQ procedure of Section 4.2, we briefly describe alternative methods — two based on matrix inversion (the `NaïveInversion` and `BlockInversion` methods, as detailed in Section 3.5), and one based exclusively on estimating columns of P — for computing mutual information in Gaussian graphs.

ColumnET

The `ColumnET` procedure uses nonstationary embedded tree estimation of specific columns of P to compute all information measures. That is to say, the marginal error variance terms are not collectively computed using the ET variant of Section 4.1.1. Given a previous selection set $\mathcal{A}$, and an ordering R over $\mathcal{R}$, the columns of $P_{\cdot|\mathcal{A}\cup R_{k-1}}$ corresponding to $\{r_k\} \cup \mathcal{S} \setminus \mathcal{A}$ are estimated via (4.3) and (4.12). The complexity of a greedy update using `ColumnET` is $\mathcal{O}(N|\mathcal{R}||\mathcal{S}|d^4)$ operations per ET iteration.

4.3.2 "Hoop-tree" Examples

To investigate the performance benefits of ET-MIQ, we consider a subclass of scalar ($d = 1$) loopy graphs containing m simple cycles (achordal "hoops") of length l, where cycles may share vertices but no two cycles may share edges. The structure of this graph resembles a macro-tree over hoop subcomponents (a "hoop-tree"; see Figure 4-1). Any embedded tree on this graph must only cut m edges ($E = m$), one for each l-cycle.

Note that the difficulty in benchmarking the performance of ET-MIQ is *not* related to the efficiency of generating a spanning tree. Rather, the difficulty lies is in generating a randomly structured loopy graph with a specified number of cycles. The class of hoop-trees is considered here primarily because it permits the generation of randomly structured loopy graphs without the subsequent need for (potentially computationally expensive) topological analysis to characterize the number of cycles and their lengths. Although ET-MIQ is applicable to any Gaussian MRF with cycles, and its complexity is dependent on the number of edges that must be cut to form an embedded tree, the class of hoop-trees affords a convenient parameterization of the number of cycles and the cycle lengths in its members – the particular randomly generated graphs.

For each problem *instance*, we generate a random hoop-tree $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of size $|\mathcal{V}| = N$. To generate a corresponding inverse covariance J, we sample $(J)_{i,j} \sim \texttt{uniform}([-1, 1])$ for each $\{i, j\} \in \mathcal{E}$, and sample $(J)_{i,i} \sim \texttt{Rayleigh}(1)$, with the

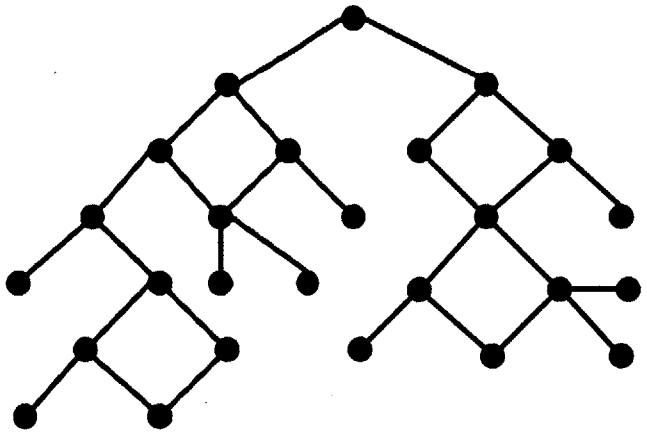


Figure 4-1: Example of a hoop-tree with 4-vertex cycles.

diagonal rescaled to enforce the positive definiteness of J. We then randomly label vertices in $\mathcal{V}$ as belonging to $\mathcal{S}$ or $\mathcal{U}$ (or neither), set a budget $\beta \propto |\mathcal{S}|$, and sample an integer-valued additive cost function $c(\cdot)$ such that $c(s) \sim \texttt{uniform}([1, \gamma\beta])$ for some $\gamma \in [0, 1]$ and all $s \in \mathcal{S}$, and such that $c(\mathcal{A}) = \sum_{a \in \mathcal{A}} c(a)$ for all $\mathcal{A} \subseteq \mathcal{S}$.

Let $\mathcal{G}_{\mathcal{T}_1}$ and $K_{\mathcal{T}_1}$ be the embedded subtree and associated regular cutting matrix formed by cutting the edge of each l-cycle with the highest absolute precision parameter $|(J)_{i,j}|$. Guided by the empirical results of [85], the second embedded tree $\mathcal{G}_{\mathcal{T}_2}$ is selected such that in every l-cycle, $K_{\mathcal{T}_2}$ cuts the edge farthest from the corresponding cut edge in the $G_{\mathcal{T}_1}$ (modulo some tie-breaking for odd l).

Figure 4-2 summarizes a comparison of ET-MIQ against `NaïveInversion`, `BlockInversion`, and `ColumnET` in terms of the mean runtime to complete a full greedy selection. Random networks of size N were generated, with $|\mathcal{R}| = 5$ and $|\mathcal{S}| = 0.3N$. The alternative methods were suppressed when they began to take prohibitively long to simulate (e.g., $N = 1200$ for `BlockInversion` and `ColumnET`).

The runtime of ET-MIQ, which vastly outperforms the alternative methods for this problem class, appears to grow superlinearly, but subquadratically, in N (approximately, bounded by $o(N^{1.7})$). The growth rate is a confluence of three factors: the $\mathcal{O}(N|\mathcal{R}|Ed^4)$ complexity per Richardson iteration of updating $\{I(\mathcal{R}; y|\mathcal{A})\}_{y \in \mathcal{S} \setminus \mathcal{A}}$; the number of Richardson iterations until the normalized residual error converges to a fixed tolerance of $\epsilon = 10^{-10}$; and the growth rate of $|\mathcal{S}|$ as a function of N, which in-

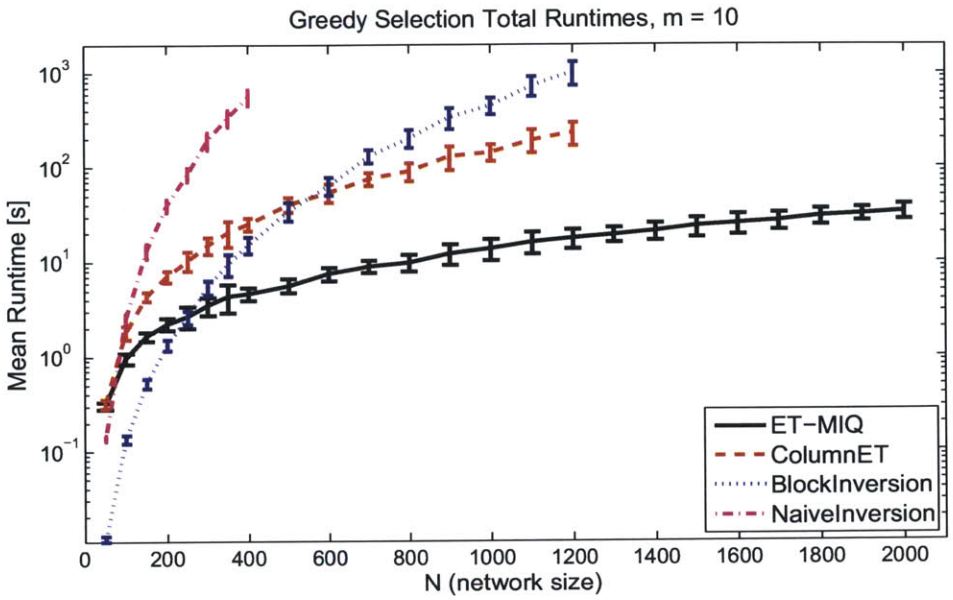


Figure 4-2: Mean runtime of the full greedy selection as a function of the network size N for randomized loopy graphs with $m = 10$ simple cycles of length $l = 4$.

directly affects the runtime through the budget β by permitting larger selection sets, and hence more rounds of greedy selection. To better disambiguate the second and third factors, we studied how the number of Richardson iterations to convergence (for a random input $\mathbf{h}$; cf. (2.18)) varies as a function of N and found no significant correlation in the case where m is constant (not a function of N). The median iteration count was 7, with standard deviation of 0.6 and range 5-9 iterations.

We also considered the effect of letting m, the number of cycles in the graph, vary with N. A runtime comparison for $m = 0.1N$ is shown in Figure 4-3. Given that $E = m = \mathcal{O}(N)$ and $|\mathcal{R}| = \mathcal{O}(1)$, ET-MIQ has an asymptotic complexity of $\mathcal{O}(N|\mathcal{R}|Ed^4) = \mathcal{O}(N^2)$. Similarly, the complexity of `ColumnET` is $\mathcal{O}(N|\mathcal{R}||\mathcal{S}|d^4) = \mathcal{O}(N^2)$. Figure 4-3 confirms this agreement of asymptotic complexity, with ET-MIQ having a lower constant factor.

We repeated the convergence study for $m = 0.1N$ and varying $N \in [100, 2000]$ (see Figure 4-4). The mean iteration count appears to grow sublinearly in N; the *actual* increase in iteration count over N is quite modest.

The relationship between the convergence and the problem structure was more

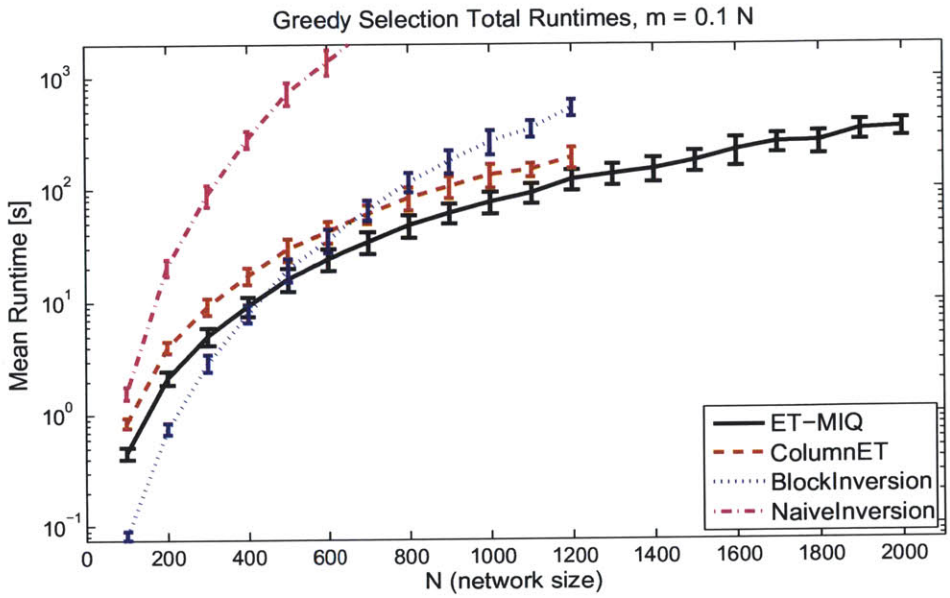


Figure 4-3: Mean runtime of the full greedy selection as a function of the network size N for randomized loopy graphs with $m = 0.1N$ simple cycles of length $l = 4$. As predicted, in the case where $m = \mathcal{O}(N)$, the ET-based algorithms have the same asymptotic complexity; ET-MIQ has a lower constant factor.

clearly illustrated when we fixed a network size of $N = 1600$ and varied the number of 4-vertex cycles $m = \delta N$, for $\delta \in [0.04, 0.32]$ (see Figure 4-5). The cycle fraction δ is strongly correlated with the iteration count – and even slightly more correlated with its log – suggesting an approximately linear (and perhaps marginally sublinear) relationship with δ, albeit with a very shallow slope.

4.4 Discussion

This chapter has presented a method of computing nonlocal mutual information in Gaussian graphical models containing both cycles and nuisances. The base computations are iterative and performed using trees embedded in the graph. We assess the proposed algorithm, ET-MIQ, and its alternatives (cf. Sections 4.1.2 and 4.3.1) in terms of the asymptotic complexity of performing a greedy update. For ET-MIQ, per-iteration complexity is $\mathcal{O}(N|\mathcal{R}|Ed^4)$, where N is the number of vertices in the

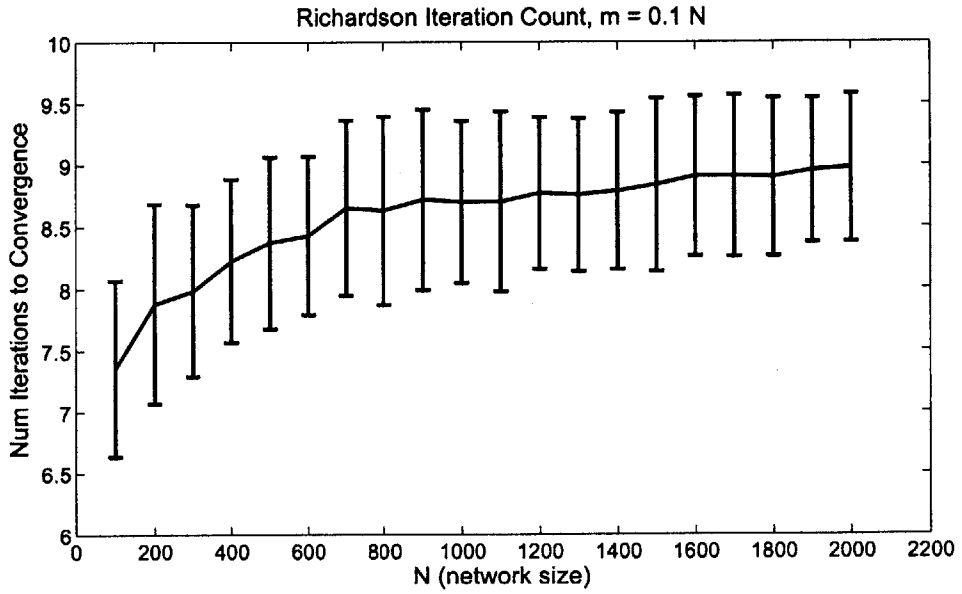


Figure 4-4: Number of Richardson iterations until convergence, for $m = 0.1N$ cycles.

network, $\mathcal{R} \subset \mathcal{V}$ is set of relevant latent variables that are of inferential interest, E is the number of edges cut to form the embedded tree, and d is the dimension of each random vector indexed by a vertex of the graph. Let κ denote the expected number of Richardson iterations to convergence of ET-MIQ, which is a direct function of the eigenproperties of the loopy precision matrix and its embedded trees and an indirect function of the other instance-specific parameters (number of cycles, network size, etc.). The experimental results of Section 4.3 suggest that the proposed algorithm, ET-MIQ, achieves significant reduction in computation over inversion-based methods, which have a total complexity of $\mathcal{O}(N^3|\mathcal{S}|d^3)$, where $\mathcal{S} \subset \mathcal{V}$ is the set of observable vertices that one has the option of selecting to later realize.

Based on the asymptotic complexities, we expect ET-MIQ would continue to achieve a significant reduction in computation for large networks whenever $|\mathcal{R}|Ed\kappa = o(N^2|\mathcal{S}|)$. Typically, the vertex dimension d is not a function of the network size. For dense networks ($|\mathcal{E}| = \mathcal{O}(N^2)$), we would not expect significant performance improvements using ET-MIQ; however, it is often the case that $\mathcal{E}$ is sparse in the sense that the number of cut edges $E = \mathcal{O}(N)$. With $|\mathcal{S}| = \mathcal{O}(N)$ (the number

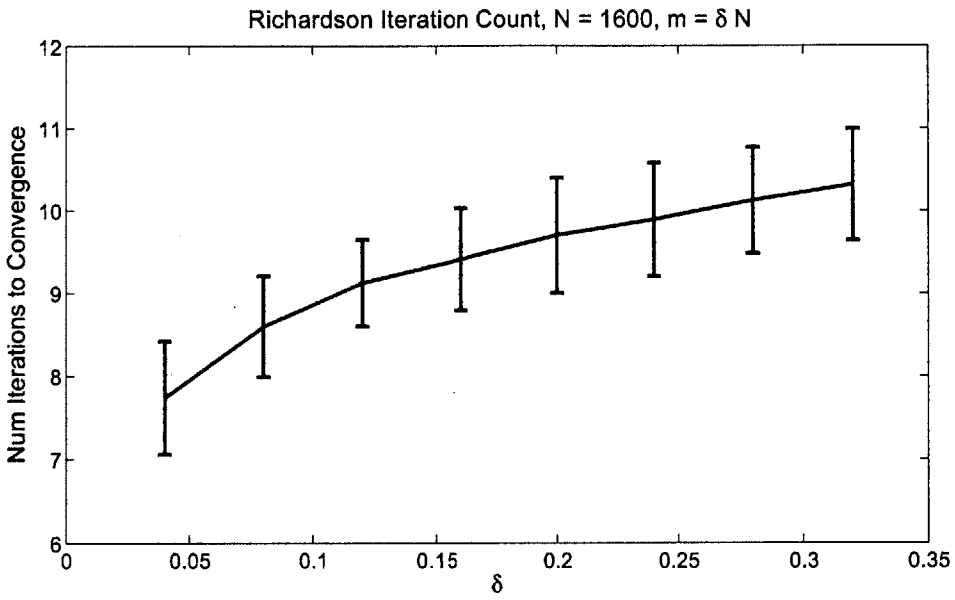


Figure 4-5: Number of Richardson iterations until convergence, for $N = 1600$ vertices, $m = \delta N$ cycles.

of available observations growing linearly in the network size), asymptotic benefits would be apparent for $|\mathcal{R}|\kappa = o(N^2)$. Since we suspect κ grows sublinearly (and very modestly) in N, and whichever system utilizing the graphical model is free to choose $\mathcal{R}$, we expect that ET-MIQ would be beneficial for efficiently quantifying information in a wide class of active inference problems on Gaussian graphs.

Chapter 5

Performance Bounds via Submodular Relaxations

Given the combinatorial nature of the focused active inference problem of Section 1.2, it is often intractable to evaluate the information objective $I(\mathcal{R}; \mathcal{A})$ for all $\mathcal{A} \in 2^{\mathcal{S}}$. Practitioners often resort to suboptimal heuristics for selecting observations; the objective then shifts to deriving appropriate bounds on the suboptimality of such heuristics.

This chapter concerns the derivation and improvement of lower bounds on performance (or, equivalently, upper bounds on the optimality gap) for the class of *general* Markov random fields, to which *arbitrary, non-Gaussian distributions with nuisances* may be Markov. Additional background material on MRF topology, submodularity, and associated performance bounds are provided in Section 5.1. An alternative performance bound predicated on the notion of a *submodular relaxation* is derived for general graphical models with nuisances in Section 5.2. Since there exist many submodular relaxations for a given focused active inference problem, the properties of *optimal* submodular relaxations – which provide the tightest such performance bound – are discussed in Section 5.3. Heuristics for approximating optimal submodular relaxations are presented in Section 5.4 and benchmarked in Section 5.5.

5.1 Preliminaries

5.1.1 More MRF Topology

For subsets $A, B, C \subset \mathcal{V}$, if $\mathcal{P}(A, B|C) = \emptyset$, then C is a *separator set* for A and B. Given that a graph can be composed of one or many connected *components*, it is useful to define functions that map vertex sets to those that are adjacent, reachable, and "retaining." Let $A \subseteq \mathcal{V}$ be a subset of vertices. Let $\Gamma : 2^{\mathcal{V}} \to 2^{\mathcal{V}}, A \mapsto \{j : i \in A, \{i, j\} \in \mathcal{E}\}$ be a *neighborhood* function that returns all vertices adjacent to (i.e., that share at least one edge with) elements of A. Furthermore, for any $C \subseteq \mathcal{V} \setminus A$, let $\Xi : 2^{\mathcal{V}} \times 2^{\mathcal{V}} \to 2^{\mathcal{V}}$ be the *reachability* function defined such that $\Xi(A|C) \triangleq \{j : \mathcal{P}(A, j|C) \neq \emptyset\}$. Finally, let $\Omega : 2^{\mathcal{V}} \times 2^{\mathcal{V}} \to 2^{\mathcal{V}}$ denote the *boundary* function, which relates a vertex set to the retaining vertices of its reachability set and is defined such that $\Omega(A|C) = \Gamma(\Xi(A|C)) \cap C$. Figure 5-1 pictorially summarizes these graph-connective functions.

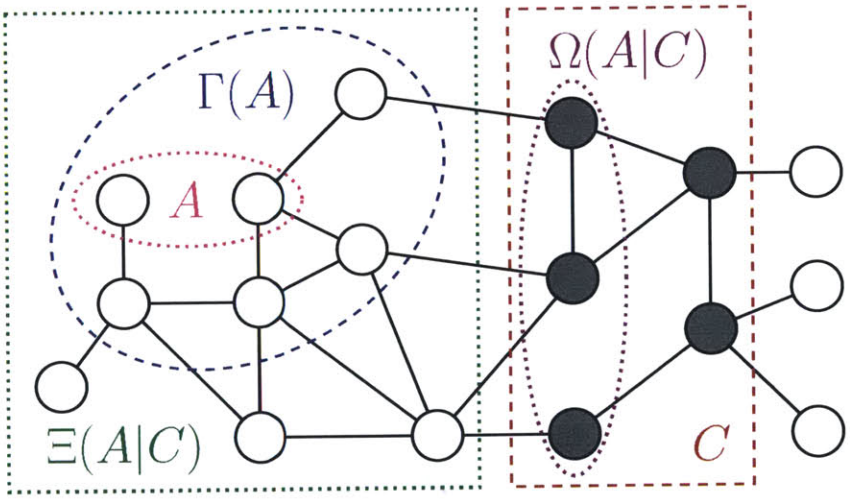


Figure 5-1: Summary of graph connectivity terms.

It is often of interest to study self-separation: instead of separating two sets, the objective is to separate every element of a particular set from all other elements.

Definition 12. For any $A \subset \mathcal{V}$, we say that $C \subseteq \mathcal{V} \setminus A$ is a *dissector* of A iff $\mathcal{P}(i, j|C) = \emptyset$ for all distinct $i, j \in A$.

Thus, if C is an A-dissector, then $\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j \mid \mathbf{x}_C$, for all distinct $i, j \in A$. The existence of a dissector is a local property relating to the nonadjacency of vertices in A.

Lemma 13. *Given an MRF $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a vertex set $A \subseteq \mathcal{V}$ is dissectible if and only if $A \cap \Gamma(A) = \emptyset$. That is, $\nexists \{i, j\} \in \mathcal{E}$ such that $i, j \in A$.*

We will hereafter denote the set of all A-dissectors as $\mathfrak{D}(A) \triangleq \{D \in 2^{\mathcal{V} \setminus A} : D \text{ dissects } A\}$.

5.1.2 Submodularity

For any set Z, a set function $F : Z \to \mathbb{R}$ is *submodular* if for every $A \subset A' \subseteq Z$, $b \notin A'$, the diminishing returns property

$$F(A \cup \{b\}) - F(A) \geq F(A' \cup \{b\}) - F(A') \tag{5.1}$$

is satisfied; that is, the marginal improvement in objective F is higher when adding an element to a particular set than any of its supersets. An equivalent, though less immediately useful, definition of submodularity would require that for all $A, B \subseteq Z$, $F(A) + F(B) \geq F(A \cup B) + F(A \cap B)$.

A set function F is monotonic if for all $A \subseteq Z$, $y \in Z$, $F(A \cup \{y\}) \geq F(A)$. The following theorem relates monotone submodular set functions to performance bounds in subset selection problems.

Theorem 14 (Nemhauser et al. [64]). *Let F be a monotone submodular set function over a finite ground set Z with $F(\emptyset) = 0$. Let A_k^g be the set of the first k elements chosen by the greedy algorithm. Then*

$$F(A_k^g) \geq \left(1 - \left(\frac{k-1}{k}\right)^k\right) \max_{\{A \subseteq Z : |A| \leq k\}} F(A) \geq \left(1 - \frac{1}{e}\right) \max_{\{A \subseteq Z : |A| \leq k\}} F(A).$$

5.1.3 Unfocused Selection Performance Bounds

Let $\mathcal{S}$ and $\mathcal{B}$ be disjoint subsets of $\mathcal{V}$, and let $F(\mathcal{A}) \triangleq I(\mathcal{B}; \mathcal{A})$. It is proven in [39] that if $\mathcal{B} \in \mathfrak{D}(\mathcal{S})$, then F is submodular and nondecreasing on $\mathcal{S} \supseteq \mathcal{A}$. If $\mathcal{S}$ is dissectible

(cf. Lemma 13), then $\mathcal{B} := \mathcal{U}$ is necessarily an $\mathcal{S}$-dissector ($\mathfrak{D}(\mathcal{S}) \neq \emptyset \Rightarrow \mathcal{U} \in \mathfrak{D}(\mathcal{S})$). Therefore, $I(\mathcal{U}; \mathcal{A})$ is submodular and nondecreasing on $\mathcal{S} \supseteq \mathcal{A}$. Thus, in the case of unit-cost observations (i.e., $c(\mathcal{A}) = |\mathcal{A}|$ for all $\mathcal{A} \subseteq \mathcal{S}$), the unfocused greedy selection $\mathcal{A}_k^g$ of cardinality k satisfies the performance bound [39]

$$I(\mathcal{U}; \mathcal{A}_k^g) \geq \left(1 - \frac{1}{e}\right) \max_{\{\mathcal{A} \subset \mathcal{S} : |\mathcal{A}| \leq k\}} I(\mathcal{U}; \mathcal{A}). \tag{5.2}$$

Remark 15. For any *proper* latent subset $\mathcal{B} \subset \mathcal{U}$ such that $\mathcal{B} \in \mathfrak{D}(\mathcal{S})$, the performance bound

$$I(\mathcal{B}; \mathcal{A}_k^g) \geq \left(1 - \frac{1}{e}\right) \max_{\{\mathcal{A} \subset \mathcal{S} : |\mathcal{A}| \leq k\}} I(\mathcal{B}; \mathcal{A}). \tag{5.3}$$

can likewise be established, although such bounds were not employed in prior literature, which has hitherto only considered unfocused selection.

5.2 Focused Selection Performance Bounds

In the *focused* variant of the active inference problem, only a proper subset $\mathcal{R} \subset \mathcal{U}$ of *relevant* latent states is of inferential interest, and nuisance variables $\mathcal{U} \setminus \mathcal{R}$ act merely as intermediaries between $\mathcal{S}$ and $\mathcal{R}$ in the joint distribution $p_{\mathbf{x}}(\cdot)$. If $\mathcal{S}$ is dissectible ($\mathfrak{D}(\mathcal{S}) \neq \emptyset$), then from Section 5.1.3, it is clear that $\mathcal{U} \in \mathfrak{D}(\mathcal{S})$. However, a primary complication of focused active inference is that $\mathcal{R} \subset \mathcal{U}$ is not necessarily an $\mathcal{S}$-dissector, so $I(\mathcal{R}; \mathcal{A})$ is not necessarily submodular. Inclusion of nuisances in the model can therefore be conceptualized as a loss of certain beneficial conditional independences, and hence a loss in the applicability of existing submodularity-based performance bounds.

Further compounding the issue is the fact that marginalizing out nuisances $\mathcal{U} \setminus \mathcal{R}$ – in addition to being potentially computationally expensive – induces a marginal graph in which $\mathcal{S}$ is may no longer be dissectible (cf. Figure 2-1).

This thesis proposes an augmentation of the relevant set $\mathcal{R}$ such that submodularity is guaranteed and relates the performance bound to this augmented set.

More concretely, an online-computable performance bound for focused observation selection problems is derived by introducing the concept of a *submodular relaxation* $\hat{\mathcal{R}} \in 2^{\mathcal{U}}$, where "relaxation" connotes that $\hat{\mathcal{R}} \supset \mathcal{R}$, and "submodular" connotes that $\hat{\mathcal{R}} \in \mathfrak{D}(\mathcal{S})$. Therefore, let $\hat{\mathfrak{R}} \triangleq \{\hat{\mathcal{R}} \in 2^{\mathcal{U}} : \hat{\mathcal{R}} \supset \mathcal{R},\ \hat{\mathcal{R}} \in \mathfrak{D}(\mathcal{S})\}$ denote the set of all feasible submodular relaxations, which is guaranteed to be nonempty whenever $\mathcal{S}$ is dissectible. By [39, Corollary 4], for any $\hat{\mathcal{R}} \in \hat{\mathfrak{R}}$, $I(\hat{\mathcal{R}}; \mathcal{A})$ is submodular and nondecreasing on $\mathcal{S} \supseteq \mathcal{A}$.

Let $\mathcal{A}^g_\beta : 2^{\mathcal{U}} \to 2^{\mathcal{S}}$ be a function mapping a latent subset $D \subseteq \mathcal{U}$ to the associated subset $\mathcal{A} \subseteq \mathcal{S}$ that greedily maximizes the focused MI measure $I(D; \mathcal{A})$, subject to budget $\beta \in \mathbb{R}_{\geq 0}$, according to the sequential greedy algorithm of Section 2.5. Assuming unit-cost observations, a greedily selected subset $\mathcal{A}^g_\beta(\hat{\mathcal{R}})$ of cardinality β satisfies the performance bound

$$I(\hat{\mathcal{R}}; \mathcal{A}^g_\beta(\hat{\mathcal{R}})) \geq \left(1 - \frac{1}{e}\right) \max_{\{\mathcal{A} \subseteq \mathcal{S} : |\mathcal{A}| \leq \beta\}} I(\hat{\mathcal{R}}; \mathcal{A}) \tag{5.4}$$

$$= \left(1 - \frac{1}{e}\right) \max_{\{\mathcal{A} \subseteq \mathcal{S} : |\mathcal{A}| \leq \beta\}} [I(\mathcal{R}; \mathcal{A}) + I(\hat{\mathcal{R}} \setminus \mathcal{R}; \mathcal{A} | \mathcal{R})] \tag{5.5}$$

$$\geq \left(1 - \frac{1}{e}\right) \max_{\{\mathcal{A} \subseteq \mathcal{S} : |\mathcal{A}| \leq \beta\}} I(\mathcal{R}; \mathcal{A}), \tag{5.6}$$

where (5.4) is due to (5.3), (5.5) to the chain rule of MI, and (5.6) to the nonnegativity of MI. Then the following proposition, which follows immediately from (5.6), provides an online-computable performance bound for *any distribution* $p_{\mathbf{x}}$ in which $\mathcal{S}$ is dissectible in the (unique [37]) minimal undirected I-map of $p_{\mathbf{x}}$.

Proposition 16. *For any set* $\hat{\mathcal{R}} \in \hat{\mathfrak{R}}$, *provided* $I(\hat{\mathcal{R}}; \mathcal{A}^g_\beta(\hat{\mathcal{R}})) > 0$ *and* $c(\mathcal{A}) = |\mathcal{A}|$ *for all* $\mathcal{A} \subseteq \mathcal{S}$, *an online-computable performance bound for any* $\bar{\mathcal{A}} \subseteq \mathcal{S}$ *in the original focused problem with relevant set* $\mathcal{R}$ *is*

$$I(\mathcal{R}; \bar{\mathcal{A}}) \geq \delta_{\mathcal{R}}(\bar{\mathcal{A}}, \hat{\mathcal{R}}) \max_{\{\mathcal{A} \subseteq \mathcal{S} : |\mathcal{A}| \leq \beta\}} I(\mathcal{R}; \mathcal{A}), \tag{5.7}$$

where

$$\delta_{\mathcal{R}}(\bar{\mathcal{A}}, \hat{\mathcal{R}}) \triangleq \left[\frac{I(\mathcal{R}; \bar{\mathcal{A}})}{I(\hat{\mathcal{R}}; \mathcal{A}^g_\beta(\hat{\mathcal{R}}))}\right]\left(1 - \frac{1}{e}\right). \tag{5.8}$$

Proposition 16 can be used at runtime to determine what percentage $\delta_{\mathcal{R}}(\bar{\mathcal{A}}, \hat{\mathcal{R}})$ of the optimal objective is guaranteed, for any focused selector, despite the lack of conditional independence of $\mathcal{S}$ conditioned on $\mathcal{R}$. Note that $\delta_{\mathcal{R}}(\bar{\mathcal{A}}, \hat{\mathcal{R}})$ need not be bounded above by $(1 - 1/e)$, as the coefficient in (5.8) may exceed unity. In order to compute the bound, a greedy heuristic running on a separate, surrogate problem with $\hat{\mathcal{R}}$ as the relevant set is required. The remainder of this chapter is devoted to the problem of finding an $\hat{\mathcal{R}} \supset \mathcal{R}$ providing the tightest bound.

5.3 Optimal Submodular Relaxations

In light of Proposition 16, it is natural to inquire if there exists an *optimal submodular relaxation*, i.e., an $\hat{\mathcal{R}} \in \hat{\mathfrak{R}}$ that provides the tightest performance bound in (5.7) by minimizing the denominator of (5.8). Given MRF $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, observable subset $\mathcal{S} \subset \mathcal{V}$ (assumed dissectible), and relevant set $\mathcal{R} \subset \mathcal{V} \setminus \mathcal{S}$, the problem statement is then

$$\underset{\hat{\mathcal{R}} \in \hat{\mathfrak{R}}}{\text{minimize}} \quad I(\hat{\mathcal{R}}; \mathcal{A}_{\beta}^{g}(\hat{\mathcal{R}})). \tag{5.9}$$

There are $2^{(|\mathcal{U}|-|\mathcal{R}|)}$ possible relaxations of $\mathcal{R}$, so the feasible set $\hat{\mathfrak{R}}$ has worst-case cardinality exponential in the number of nuisance variables. Therefore, confirmation that a submodular relaxation is optimal in the sense of (5.9) can have exponential-time complexity, which may be prohibitive for large graphs (i.e., high-dimensional distributions) with many nuisance variables. Marginalization of nuisances lying outside the optimal submodular relaxation, in an effort to truncate the search space over $\hat{\mathfrak{R}}$, remains impracticable because specifying such an elimination order presupposes knowledge of the members of the optimal submodular relaxation.

Several properties of submodular relaxations can be elicited by considering the following example.

Example 2. Let $D, D' \in \hat{\mathfrak{R}}$ be submodular relaxations of $\mathcal{R}$ that are strictly nested

(in the sense that $D \subset D'$), and consider without loss of generality[1] the case where $D' \setminus D = \{\tilde{d}\}$.

For any $\mathcal{A} \subset \mathcal{S}$, and for each $s \in \mathcal{S} \setminus \mathcal{A}$,

$$\begin{aligned} I(s; D'|\mathcal{A}) &\overset{(a)}{=} I(s; D|\mathcal{A}) + I(s; \tilde{d} \mid D \cup \mathcal{A}) \\ &\overset{(b)}{=} I(s; D|\mathcal{A}) + I(s; \tilde{d}|D), \end{aligned}$$

where (a) is due to the chain rule of MI, and (b) is due to $D \in \hat{\mathfrak{R}}$, whereby $\mathcal{P}(s, \mathcal{A}|D) = \emptyset$ for any $s \in \mathcal{S} \setminus \mathcal{A}$. The greedy selection $\mathcal{A}_\beta^g(D)$ is formed by selecting at step $k \in \{1, \ldots, \beta\}$

$$\begin{aligned} s_k &\in \underset{s \in \mathcal{S} \setminus \mathcal{A}_{k-1}}{\operatorname{argmax}} I(s; D|\mathcal{A}_{k-1}) \\ \mathcal{A}_k &\leftarrow \mathcal{A}_{k-1} \cup \{s_k\}, \end{aligned} \tag{5.10}$$

while for $\mathcal{A}_\beta^g(D')$,

$$\begin{aligned} s'_k &\in \underset{s \in \mathcal{S} \setminus \mathcal{A}'_{k-1}}{\operatorname{argmax}} I(s; D|\mathcal{A}'_{k-1}) + I(s; \tilde{d}|D) \\ \mathcal{A}'_k &\leftarrow \mathcal{A}'_{k-1} \cup \{s'_k\}. \end{aligned} \tag{5.11}$$

Since $D \in \hat{\mathfrak{R}}$, then for the graph conditioned on D, $\tilde{d}$ lies in a component containing at most one $s \in \mathcal{S}$. Thus, $I(s; \tilde{d}|D) > 0$ for at most one $s \in \mathcal{S}$. The effect of removing $\tilde{d}$ from $D' \in \hat{\mathfrak{R}}$ can be better delineated by considering two cases for where $\tilde{d}$ lies.

Case 1: $\tilde{d} \notin \Xi\left(\mathcal{A}_\beta^g(D') \mid D\right)$. In this case, $I(s; \tilde{d}|D) = 0$ for all $s \in \mathcal{A}_\beta^g(D')$. Therefore, by comparison of (5.11) with (5.10), $\{s_k\}_{k=1}^\beta = \{s'_k\}_{k=1}^\beta \Rightarrow \mathcal{A}_\beta^g(D) = \mathcal{A}_\beta^g(D')$. Thus,

$$\begin{aligned} I(D'; \mathcal{A}_\beta^g(D')) &\overset{(a)}{=} I(D; \mathcal{A}_\beta^g(D')) + I(\tilde{d}; \mathcal{A}_\beta^g(D')|D) \\ &\overset{(b)}{=} I(D; \mathcal{A}_\beta^g(D')) \\ &\overset{(c)}{=} I(D; \mathcal{A}_\beta^g(D)), \end{aligned}$$

[1]The general case where $|D' \setminus D| \geq 1$ can be established through induction.

where (a) is due to the chain rule of MI, (b) to the assumption that $\tilde{d} \notin \Xi\left(\mathcal{A}_\beta^g(D') \mid D\right)$, and (c) to the result above.

Case 2: $\tilde{d} \in \Xi\left(\mathcal{A}_\beta^g(D') \mid D\right)$. There exists exactly one $\tilde{s} \in \mathcal{A}_\beta^g(D')$ such that $\tilde{d} \in \Xi(\tilde{s}|D)$, and $\mathcal{P}(\tilde{d}, s|D) = \emptyset$ for all $s \in \mathcal{S} \setminus \{\tilde{s}\}$. Suppose $\tilde{s}$ is chosen at step $\tilde{k} \in \{1, \ldots, \beta\}$ of (5.11). Then $\mathcal{A}_{\tilde{k}-1}^g(D) = \mathcal{A}_{\tilde{k}-1}^g(D')$.

In the (possible) case where the deletion of $\tilde{d}$ from D' does not change the greedy selection such that $\mathcal{A}_\beta^g(D) = \mathcal{A}_\beta^g(D')$, then the reduction in the objective $I(\cdot, \mathcal{A}_\beta^g(\cdot))$ is exactly $I(D'; \mathcal{A}_\beta^g(D')) - I(D; \mathcal{A}_\beta^g(D)) = I(D'; \mathcal{A}_\beta^g(D)) - I(D; \mathcal{A}_\beta^g(D)) = I(\tilde{d}; \tilde{s}|D)$.

However, for $k = \tilde{k}, \ldots, \beta$, $\mathcal{A}_k^g(D)$ and $\mathcal{A}_k^g(D')$ may diverge due to the sensitivity of the greedy algorithm to upstream changes in the incremental marginal value of a selection. In such cases, it is possible that $\mathcal{A}_\beta^g(D)$ is such that $I(D; \mathcal{A}_\beta^g(D)) \geq I(D'; \mathcal{A}_\beta^g(D'))$, i.e., the deletion of $\tilde{d}$ from D has led to an increase in the objective and, thus, a looser bound.

Let $\hat{\mathfrak{R}}^* \subseteq \hat{\mathfrak{R}}$ denote the set of all submodular relaxations of $\mathcal{R}$ that are optimal w.r.t. (5.9).

Proposition 17 (Existence and Nonuniqueness). *If $\mathcal{S}$ is dissectible, then $|\hat{\mathfrak{R}}^*| \geq 1$, i.e., an optimal submodular relaxation exists and is generally not unique.*

Proof. If $\mathfrak{D}(\mathcal{S}) \neq \emptyset$, then $\hat{\mathfrak{R}} \neq \emptyset$, whereby $\hat{\mathfrak{R}}^* \neq \emptyset$. It suffices to prove that there can exist cases in which $|\hat{\mathfrak{R}}^*| > 1$. Case 1 of Example 2 provides such an example: If $D \in \hat{\mathfrak{R}}^*$, then $D' \in \hat{\mathfrak{R}}^*$, so the optimal submodular relaxation is generally not unique. □

Definition 18 (Minimum-Cardinality Subset). A subset $A \subseteq C$ satisfying property Υ is a *minimum-cardinality* subset if there does not exist another subset $B \subset C$ that has lower cardinality (i.e., $|B| < |A|$) and also satisfies Υ.

Proposition 19 (Nonminimum Cardinality). *Minimum-cardinality extensions of $\mathcal{R}$ that dissect $\mathcal{S}$ are not necessarily optimal submodular relaxations.*

Proof. Consider the counterexample in Figure 5-2, in which $\mathcal{S} = \{1, 6\}, \mathcal{U} = \{2, 3, 4, 5\}$, and $\mathcal{R} = \{4\}$. Depending on the parameterization of the distribution $p_{\mathbf{x}}(\cdot)$, whose

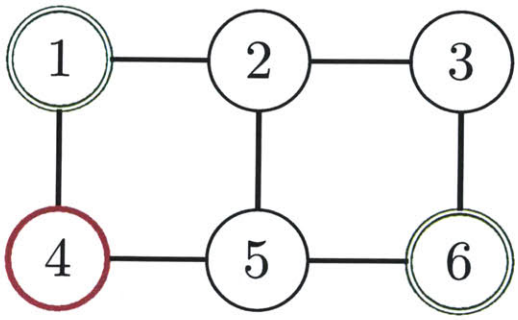


Figure 5-2: Six-vertex counterexample to minimum-cardinality, with $\mathcal{S} = \{1, 6\}$ and $\mathcal{R} = \{4\}$.

structure the graph represents, the unique optimal relaxation may be either $\hat{\mathcal{R}}^* = \{2, 4\}$ or $\hat{\mathcal{R}}^* = \{3, 4, 5\}$, the latter being of nonminimum cardinality. □

In addition to minimum-cardinality subsets, one could consider *minimal* subsets.

Definition 20 (Minimal Subset)**.** A set A satisfying property Υ is *minimal* if no proper subset $B \subset A$ satisfies Υ as well.

Let $\hat{\mathfrak{R}}_{\min} \triangleq \{\hat{\mathcal{R}} \in \hat{\mathfrak{R}} : \hat{\mathcal{R}} \text{ is minimal}\}$ denote the set of minimal feasible submodular relaxations. Furthermore, let $\hat{\mathfrak{R}}^*_{\min} \triangleq \hat{\mathfrak{R}}_{\min} \cap \hat{\mathfrak{R}}^*$. Note that for any $D \in \hat{\mathfrak{R}}$, the MI objective $I(D; \mathcal{A})$ is monotonic in $\mathcal{S} \supseteq \mathcal{A}$, i.e., in the sense that $I(D; \mathcal{A}) \leq I(D; \mathcal{A}')$ for any $\mathcal{A} \subseteq \mathcal{A}' \subseteq \mathcal{S}$. However, the function $\nu_\beta(D) \triangleq I(D; \mathcal{A}^g_\beta(D))$ is not necessarily monotonic in $\mathcal{U} \supseteq D$, as the following proposition illustrates.

Proposition 21 (Nonminimality)**.** $\hat{\mathfrak{R}}^*_{\min} = \emptyset \not\Rightarrow \hat{\mathfrak{R}} = \emptyset$*, i.e., optimal submodular relaxations are not necessarily minimal subsets.*

Proof. It suffices to provide a counterexample to the following monotonicity argument: For $D, D' \in \hat{\mathfrak{R}}$ such that $D \subseteq D'$, $I(D; \mathcal{A}^g_\beta(D)) \leq I(D'; \mathcal{A}^g_\beta(D'))$. Case 2 of Example 2 provides such a counterexample. □

In summary, optimal submodular relaxations are generally neither unique nor minimal, thus complicating the search for a submodular relaxation providing the tightest online-computable performance bounds. Furthermore, the notion of a submodular relaxation requires a particular form of vertex cover, and there do not appear to be any analogies between optimizing the choice of submodular relaxation and solu-

tions of max-flow/min-cut (i.e., edge cutting) dual problems considered in network optimization [1].

5.4 Iterative Tightening

Given the complexity of identifying optimal submodular relaxations, this section presents two heuristic methods — one direct, and one somewhat more indirect — for efficiently generating submodular relaxations with low suboptimality.

The algorithms presented in this section are predicated on the intuition that despite Proposition 21, the "best" submodular relaxations $\hat{\mathcal{R}} \supset \mathcal{R}$ are likely those that remain as close as possible to the original relevant latent set $\mathcal{R}$ while not grossly inflating the apparent informativeness of observations with respect to the augmenting nuisances $\hat{\mathcal{R}} \setminus \mathcal{R}$.

Both heuristics are based on the idea of starting with a feasible submodular relaxation $D' \in \hat{\mathfrak{R}}$ and "tightening" it by removing a single element $d \in D'$. In order to ensure that the resulting set $D \coloneqq D' \setminus \{d\}$ is also feasible, a list of *critical boundary vertices* Ω_{crit} must be maintained. The removal of any vertex from Ω_{crit} would result in a graph in which $\mathcal{S}$ is not dissected. An algorithm for determining Ω_{crit} for any tuple $\langle \mathcal{G}, \mathcal{S}, \hat{\mathcal{R}} \rangle$ is presented in Algorithm 5.

Algorithm 5 GETCRITICALBOUNDARY$(\mathcal{G}, \mathcal{S}, \hat{\mathcal{R}})$

Require: $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\mathcal{S} \subset \mathcal{V}$, $\hat{\mathcal{R}} \in \hat{\mathfrak{R}}$:
  $\Omega_{\text{crit}} \leftarrow \emptyset, \Omega_{\text{all}} \leftarrow \emptyset$
  for $s \in \mathcal{S}$ **do**
    $\Omega_{\text{crit}} \leftarrow \Omega_{\text{crit}} \cup \left(\Omega(s|\hat{\mathcal{R}}) \cap \Omega_{\text{all}}\right)$
    $\Omega_{\text{all}} \leftarrow \Omega_{\text{all}} \cup \Omega(s|\hat{\mathcal{R}})$
5: **end for**
  return Ω_{crit}

In addition to critical boundary vertices, the heuristic class described in this section tracks the set of observation vertices "activated" by the greedy heuristic. Speaking informally, the goal of iterative tightening is to create "space" around these activated observation vertices by removing boundary nodes from their reachable sets,

thereby reducing the information contribution of reachable augmenting nuisances and tightening the performance bound (5.7).

5.4.1 Direct

Starting from an initial solution[2] $D \in \hat{\mathfrak{R}}$, the *direct* version of ITERATIVETIGHTEN (Algorithm 6) performs a sequential-greedy minimization of $\nu_\beta(D) \triangleq I(D; \mathcal{A}_\beta^g(D))$ by considering the element $d \in D$ whose removal (constrained so that the resulting set is still a submodular relaxation) yields the lowest objective $I\left(D \setminus \{d\}; \mathcal{A}_\beta^g(D \setminus \{d\})\right)$ for the subsequent iterate. In this sense, ITERATIVETIGHTEN-DIRECT anticipates how the removal of an element affects the resulting greedy selection, but at the expense of having to determine the greedy selection $\mathcal{A}_\beta^g(D \setminus \{d\})$ for every candidate d.

Algorithm 6 ITERATIVETIGHTEN-DIRECT$(p, \mathcal{G}, \mathcal{S}, \mathcal{R}, D)$

Require: $p_{\mathbf{x}}(\cdot)$, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\mathcal{S} \subset \mathcal{V}$, $\mathcal{R} \subset \mathcal{V} \setminus \mathcal{S}$, $D \in \hat{\mathfrak{R}}$:
  $Q \leftarrow D \setminus \mathcal{R}$
  $Q \leftarrow Q \setminus \Omega_{\text{crit}}(\mathcal{G}, \mathcal{S}, D)$
  while $Q \neq \emptyset$ **do**
    $d_{\text{del}} \in \operatorname{argmin}_{d \in Q} I\left(D \setminus \{d\}; \mathcal{A}_\beta^g\left(D \setminus \{d\}\right)\right)$
    $D \leftarrow D \setminus \{d_{\text{del}}\}$
    $Q \leftarrow Q \setminus \Omega_{\text{crit}}(\mathcal{G}, \mathcal{S}, D)$
  end while
  return $\hat{\mathcal{R}} := D$

5.4.2 Indirect

Given the expense of computing a set of candidate greedy selections *at each iteration* of ITERATIVETIGHTEN-DIRECT, the indirect form of ITERATIVETIGHTEN (Algorithm 7) instead approximates the deletion of an index from the submodular relaxation by relating the deleted term to an anticipation of the objective reduction.

Starting from a submodular relaxation $D' \in \hat{\mathfrak{R}}$, suppose that by removing a $\tilde{d} \in D'$, the resulting set $D := D' \setminus \{\tilde{d}\} \in \hat{\mathfrak{R}}$. Example 2 describes such a situation. If $\mathcal{A}_\beta^g(D) = \mathcal{A}_\beta^g(D')$, then the objective function is reduced by $\max_{s \in \mathcal{S}} I(s; \tilde{d}|D)$, where

[2] In the absence of an initial solution, the trivial submodular relaxation $D := \mathcal{U} \in \hat{\mathfrak{R}}$ may be used.

Algorithm 7 IterativeTighten-Indirect$(p, \mathcal{G}, \mathcal{S}, \mathcal{R}, D)$

Require: $p_{\mathbf{x}}(\cdot)$, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\mathcal{S} \subset \mathcal{V}$, $\mathcal{R} \subset \mathcal{V} \setminus \mathcal{S}$, $D \in \hat{\mathfrak{R}}$:
  $Q \leftarrow D \setminus \mathcal{R}$
  $Q \leftarrow Q \setminus \Omega_{\text{crit}}(\mathcal{G}, \mathcal{S}, D)$
  while $Q \neq \emptyset$ **do**
    $A \leftarrow \mathcal{A}^g_\beta(D)$
    if $\Omega(A|D) \cap Q = \emptyset$ **then**
      $A \leftarrow \mathcal{S}$
    end if
    $d_{\text{del}} \in \text{argmax}_{s \in A, d \in Q}\ I(s; d \mid D \setminus \{d\})$
    $D \leftarrow D \setminus \{d_{\text{del}}\}$
    $Q \leftarrow Q \setminus \Omega_{\text{crit}}(\mathcal{G}, \mathcal{S}, D)$
  end while
  return $\hat{\mathcal{R}} := D$

$I(s; \tilde{d}|D)$ is positive for at most one $s \in \mathcal{S}$. By inspection of (5.11), it is clear that $\tilde{d}$ affects the baseline – and not marginal – value of such an s. It is in light of this distinction that $I(s; \tilde{d}|D)$ is used as an approximation of the objective decrease under small perturbations of the greedy selection.

5.5 Experiments

In this section, the two variants of IterativeTighten are benchmarked against alternative heuristics for determining submodular relaxations.

5.5.1 Heuristics for Comparison

RandConstruct (Algorithm 8) starts with the relevant latent set $\mathcal{R}$ and augments it by randomly adding latent indices until the resulting set is an $\mathcal{S}$-dissector. Conversely, RandDeconstruct (Algorithm 9) starts with any submodular relaxation and randomly removes indices until the resulting submodular relaxation is minimal w.r.t $\hat{\mathfrak{R}}$. The trivial relaxation $\mathcal{U} \in \hat{\mathfrak{R}}$ can be compactly represented by the Picket-Fence heuristic, which uses $\hat{\mathcal{R}} := \mathcal{R} \cup \Gamma(\mathcal{S})$. For networks small enough to permit enumeration of all submodular relaxations, MinCard samples uniformly from the subset of those with minimum cardinality.

Algorithm 8 RANDCONSTRUCT($\mathcal{G}, \mathcal{S}, \mathcal{R}$)

$D \leftarrow \mathcal{R}$
while $D \notin \hat{\mathfrak{R}}$ **do**
  $d_{\text{add}} \sim \texttt{unif}(\mathcal{U} \setminus D)$
  $D \leftarrow D \cup \{d_{\text{add}}\}$
end while
return $\hat{\mathcal{R}} \coloneqq D$

Algorithm 9 RANDDECONSTRUCT($\mathcal{G}, \mathcal{S}, \mathcal{R}, D$)

$Q \leftarrow D \setminus \mathcal{R}$
$Q \leftarrow Q \setminus \Omega_{\text{crit}}(\mathcal{G}, \mathcal{S}, D)$
while $Q \neq \emptyset$ **do**
  $d_{\text{del}} \sim \texttt{unif}(Q)$
  $D \leftarrow D \setminus \{d_{\text{del}}\}$
  $Q \leftarrow Q \setminus \Omega_{\text{crit}}(\mathcal{G}, \mathcal{S}, D)$
end while
return $\hat{\mathcal{R}} \coloneqq D$

5.5.2 Numerical Results (4×4)

The ITERATIVETIGHTEN heuristics of Section 5.4 were benchmarked against those of Section 5.5.1 over the course of 100 trials (see Figure 5-3). For each trial, a Gaussian prior Markov to a 4×4 nearest neighbors grid was instantiated and tested. An exhaustive search of $\hat{\mathfrak{R}}$ was tractable for this simple problem and was used to evaluate the suboptimality of all tested heuristics; during the course of the exhaustive search, a set of minimum cardinality submodular relaxation was retained, from which the MINCARD heuristic submodular relaxation was sampled uniformly.

The results of this trial are encouraging, insofar as the ITERATIVETIGHTEN heuristics provide a clear relative advantage in terms of suboptimality over the other heuristics tested. That ITERATIVETIGHTEN-INDIRECT did slightly *better* than its direct counterpart is mildly surprising, although this can easily be explained by the prevalence of local optima in the problem. The marked suboptimality of the PICKETFENCE heuristic is not surprising: By augmenting the relevant set with nuisances directly adjacent to observable vertices, it contradicts the main intuition behind ITERATIVETIGHTEN, which is of creating space between active observations and augmenting nuisances.

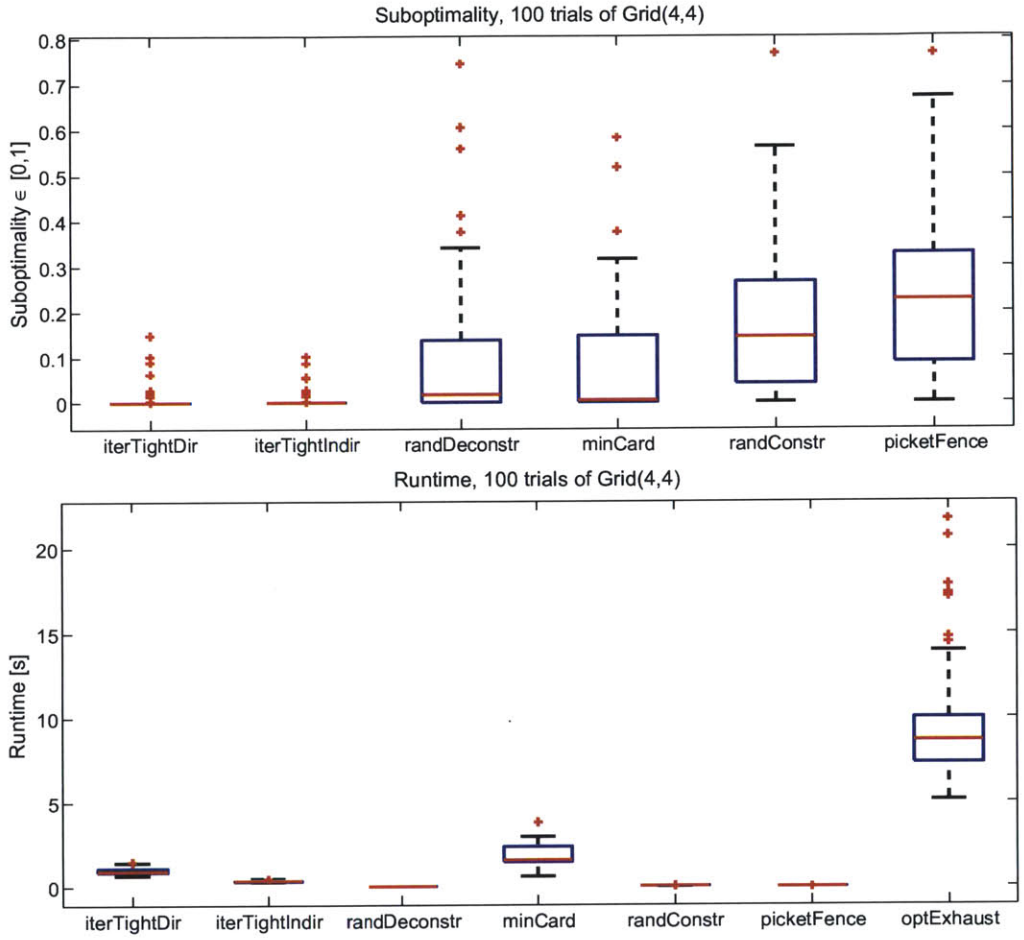


Figure 5-3: Benchmark of submodular relaxation heuristics over 100 randomly generated Gaussian distributions Markov to the 4×4 nearest neighbors grid. Interquartile ranges are shown in blue, medians in red (line segment), and outliers as $+$. Top: suboptimality as measured against $\hat{\mathfrak{R}}^*$. Bottom: Runtime in seconds, along with the runtime needed to verify the optimality via exhaustive search.

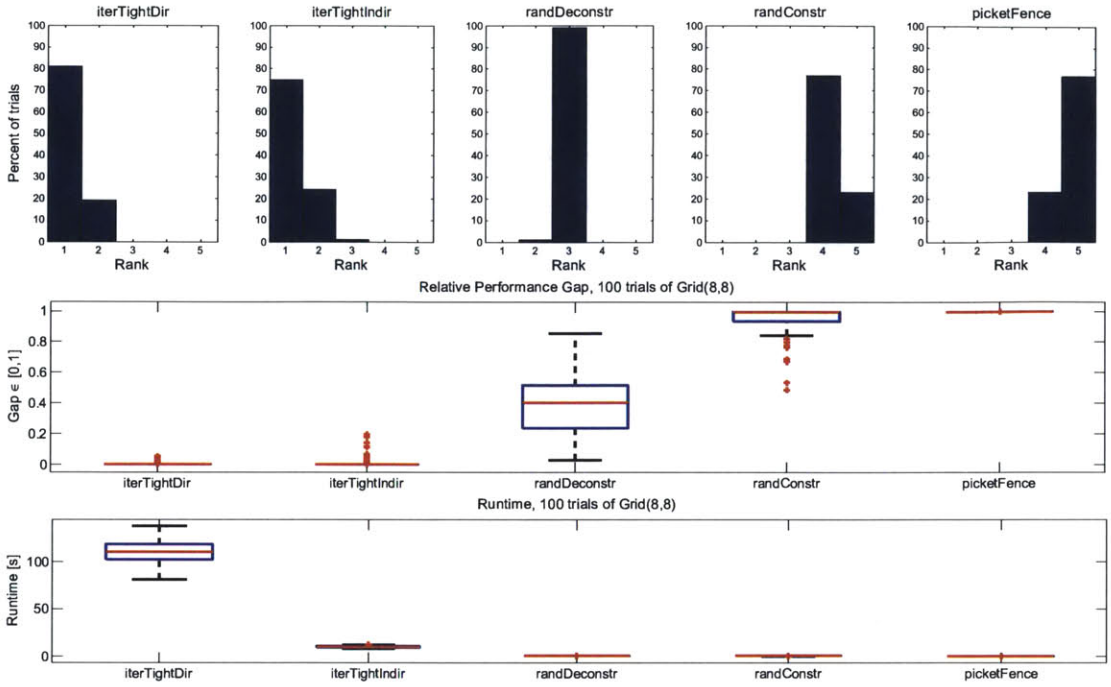


Figure 5-4: Benchmark of submodular relaxation heuristics over 100 randomly generated Gaussian distributions Markov to the 8×8 nearest neighbors grid. Top: Histogram of relative rank for the five heuristics benchmarked. Middle: Relative performance gap as measured vs. the highest scoring (lowest MI) heuristic for each of the 100 randomly generated networks. Bottom: Runtime in seconds.

5.5.3 Numerical Results (8×8)

The ITERATIVETIGHTEN heuristics of Section 5.4 were again benchmarked against those of Section 5.5.1, over 100 trials, each of which consisting of a randomly generated Gaussian distribution Markov to an 8×8 nearest neighbors grid. This network of 64 nodes was already too large to permit exhaustive enumeration and characterization of all submodular relaxations. As a consequence, the suboptimal heuristics were for each trial ranked, and the rank and spread (difference between the objective and that of the top ranked heuristic) were tallied (cf. Figure 5-4). Once again, the ITERATIVETIGHTEN heuristics show a marked improvement in objective over their random/uninformed competitors, with ITERATIVETIGHTEN-INDIRECT providing a tradeoff between performance and runtime.

5.6 Summary

This chapter has considered the class of general Markov random fields — arbitrary distributions whose conditional independence/factorization structures can be represented by undirected graphical models — and has derived performance bounds on the suboptimality of heuristic observation selectors in the presence of nuisances. The conditional independence requirements that guarantee submodularity of the focus objective, and thus applicability of existing performance bounds, are not always satisfied by the relevant latent set to be inferred. This chapter has introduced the notion of submodular relaxations as surrogate problems that can be used to establish online-computable performance bounds on the percentage of the optimal objective guaranteed by a heuristic selector. This percentage is not constrained to lie below $(1 - 1/e)$, although it must be computed online and is not known *a priori*.

Given the multitude of submodular relaxations for any particular focused active inference problem, this chapter has also investigated the *quality* of submodular relaxations. Between the original focused problem and the surrogate problem, which augments the relevant set with additional nuisances, the following intuition was offered: The submodular relaxation whose augmenting nuisances provide the least inflation in the informativeness of greedily selected observations is, in general, preferred. However, the predication of the online-computable bound on the greedy heuristic for the surrogate problem led to issues in characterizing *optimal* submodular relaxations, which provide the tightest such bounds. It was shown that optimal submodular relaxations are in general not unique, not minimum-cardinality, and not minimal. Heuristics that iteratively tighten a submodular relaxation by deleting augmenting nuisances with the strongest apparent information contributions, as per the intuition of the chapter, were shown to outperform other heuristics based on cardinality or random augmentation/deaugmentation.

Performance bounds *not* predicated on the greedy heuristic, and thus potentially more amenable to analysis, are the subject of future work.

Chapter 6

Discussion

6.1 Summary of Thesis Contributions

This thesis has addressed some of the fundamental issues surrounding efficient model-based observation selection in the presence of nuisances. Chapter 3 presented decompositions of nonlocal mutual information on universally embedded paths in Gaussian graphical models; these decompositions enabled a reduction in the complexity of greedy selection updates on Gaussian trees from cubic to linear in the network size. Chapter 4 presented a framework for efficient information quantifications in loopy Gaussian graphs by estimating particular conditional covariance matrices via embedded substructure computations. Chapter 5 introduced the concept of *submodular relaxations* to derive new online-computable bounds on suboptimality of observation selections in the presence of nuisance variables in the graph, which generally invalidate prior submodular bounds. A subsequent characterization of optimal submodular relaxations, which provide the tightest such performance bounds, yielded efficient heuristics for approximating the optimum.

6.2 Limitations and Future Work

Despite the above summarized contributions, focused active inference is far from a "solved" problem. A number of extensions to the contributions of this thesis are

possible, apparent, and would greatly contribute to the understanding of information-theoretic planning. Some of these extensions are now described.

6.2.1 Inferring Relevance

This thesis has assumed that the relevant latent set $\mathcal{R} \subseteq \mathcal{U}$ has been specified. This is not a wholly unreasonable assumption, especially when queries are posed by domain experts with experiential priors on what states are most important to infer. However, it is not hard to envision settings in which those making use of automated observation selection may have inexact preferences for what data to present to them, in which case one may attempt to infer relevant latent sets from these sequential interactions.

6.2.2 Automated Transcription

This thesis has also assumed a fully parameterized graphical model has been submitted to be analyzed. One extension that would make the contributions of this thesis more readily integrable in actual observation networks would be a tool that automates graphical model transcription. For example, a network of heterogeneous sensing resources located in some set of space-time configuration could be characterized both in terms of individual components (such as sensor modalities/noise) as well as environmental (e.g., geochemical) properties that interrelate phenomena being observed by all agents.

6.2.3 Prioritized Approximation

The methods described in this thesis are exact in the sense that all mutual information measures are estimated to within a specified tolerance. If the computational cost of quantifying mutual information were constrained (e.g., in a distributed estimation framework with communication costs), it may be of interest to develop algorithms for allowing prioritized approximation depending on how sensitive the overall information reward is to these conditional mutual information terms. In addition to algorithms for adaptively selecting embedded trees to hasten convergence, [14] propose methods for

choosing and updating only a subset of variables in each Richardson iteration. If, in an essentially dual problem to (1.1), the cost of sensor selections were to be minimized subject to a quota constraint on the minimum amount of collected information — i.e.,

$$\begin{aligned} \text{minimize}_{\mathcal{A}\subseteq\mathcal{S}} \quad & c(\mathcal{A}) \\ \text{s.t.} \quad & I(\mathbf{x}_\mathcal{R}; \mathbf{x}_\mathcal{A}) \geq \alpha, \end{aligned} \tag{6.1}$$

where $\alpha \in \mathbb{R}_{\geq 0}$ is an information quota that must be achieved — the ability to truncate information quantification when a subset of the graph falls below an informativeness threshold would be of potential interest.

6.2.4 Weighted Mutual Information

The objective function considered in this thesis is focused mutual information $I(\mathcal{R}; \mathcal{A})$, where the relevant latent set $\mathcal{R} \subset \mathcal{U}$ is provided a priori and the observations $\mathcal{A} \subseteq \mathcal{S}$ must be chosen to maximize the objective (subject to budget constraints). Generally, even if several or many variables are deemed relevant, their relevancy might not be uniform, in which case one may wish to assign relative weights to the individual variables. Weighted mutual information was devised for such a purpose [29], although it can be shown in some cases to be negative [44]. Although there have been recent attempts to patch weighted MI to enforce nonnegativity [71], it remains to be seen if weighted MI possesses decompositions similar to those of its unweighted counterparts.

6.2.5 Non-Gaussian Distributions

The contributions of this thesis aimed at efficient information quantification have assumed Gaussianity. While multivariate Gaussian distributions are often reasonable approximations for real-valued unimodal data, there are a variety of domains with truncated support or multiple underlying modes. A number of impediments to moving beyond the Gaussian regime persist. Models as simple as Gaussian mixtures lack closed-form solutions for mutual information. Few classes of continuous distributions

are closed under marginalization and conditioning. Moments beyond the second moment may need to be exchanged to update beliefs. Despite recent developments to extend continuous variable belief propagation beyond multivariate Gaussianity (e.g., through particle BP [32], nonparametric BP [84], kernel BP [81], and nonparanormal BP [22]), information quantification in non-Gaussian regimes remains a worthwhile, albeit formidable, challenge.

6.2.6 Other f-divergence Information Measures

Mutual information is but one (arguably fundamental) information measure specified as a particular evaluation of relative entropy (2.8). However, there exists a more general class of information measures specified by f-divergences [2, 20] of the form

$$D_f(p \,\|\, q) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)\, dx,$$

where f is a convex function. For example, $f(t) := t \log t$ recovers the relative entropy measure. Due to the connection between f-divergences and surrogate loss functions [6, 65], extensions of the contributions of this thesis to other information measures would further the understanding of efficient data acquisition and processing in other statistical settings.

Bibliography

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications.* Prentice Hall, 1 edition, February 1993.

[2] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B (Methodological)*, 28(1):131–142, 1966.

[3] B. Anderson and A. Moore. Active learning for hidden Markov models: Objective functions and algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2005.

[4] B. S. Anderson and A. W. Moore. Fast information value for graphical models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 18, 2005.

[5] A. C. Atkinson and A. N. Donev. *Optimal Experimental Designs.* Oxford University Press, 1992.

[6] P. L. Bartlett, M. I. Jordan, and McAullife. Convexity, classification, and risk bounds. *Journal of the American Statistical Society*, 101:138–156, 2006.

[7] K. A. C. Baumgartner, S. Ferrari, and A. V. Rao. Optimal control of an underwater sensor network for cooperative target tracking. *IEEE Journal of Oceanic Engineering*, 34(4):678–697, 2009.

[8] B. Bonet and H. Geffner. Planning with incomplete information as a heuristic search in belief space. In *Proceedings of the Artificial Intelligence Planning Systems Conference*, 2000.

[9] A. Bry and N. Roy. Rapidly-exploring random belief trees for motion planning under uncertainty. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 2011.

[10] W. F. Caselton and J V. Zidek. Optimal monitoring network designs. *Statistics and Probability Letters*, 2(4):223–227, 1984.

[11] A. Cassandra, L. Kaelbling, and J. A. Kurien. Acting under uncertainty: Discrete Bayesian models for mobile-robot navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1996.

[12] D. A. Castanon. Approximate dynamic programming for sensor management. In *Proceedings of the IEEE Conference on Decision and Control (CDC)*, volume 2, pages 1202–1207. IEEE, December 1997.

[13] CERN - European Organization for Nuclear Research. Colt, 1999.

[14] V. Chandrasekaran, J. K. Johnson, and A. S. Willsky. Estimation in Gaussian graphical models using tractable subgraphs: A walk-sum analysis. *IEEE Transactions on Signal Processing*, 56(5):1916–1930, May 2008.

[15] H.-L. Choi. *Adaptive sampling and forecasting with mobile sensor networks.* PhD thesis, Massachusetts Institute of Technology, 2008.

[16] H.-L. Choi and J. P. How. Continuous trajectory planning of mobile sensors for informative forecasting. *Automatica*, 46(8):1266–1275, 2010.

[17] M. J. Choi, V. Y. F. Tan, A. Anandkumar, and A. S. Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research (JMLR)*, 12:1771–1812, May 2011.

[18] D. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research (JAIR)*, 4:129–145, 1996.

[19] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, second edition, 2006.

[20] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[21] V. Delouille, R. Neelamani, and R. G. Baraniuk. Robust distributed estimation using the embedded subgraphs algorithm. *IEEE Transactions on Signal Processing*, 54:2998–3010, 2006.

[22] G. Elidan and C. Cario. Nonparanormal belief propagation (NPNBP). In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, 2012.

[23] P. L. Erdős, M. A. Steel, L. A. Székely, and T. J. Warnow. A few logs suffice to build (almost) all trees: Part ii. *Theoretical Computer Science*, 221:77–118, 1999.

[24] H. J. S. Feder, J. J. Leonard, and C. M. Smith. Adaptive mobile robot navigation and mapping. *International Journal of Robotics Research (IJRR)*, 18(7):650–668, July 1999.

[25] V. V. Fedorov. *Theory of optimal experiments.* Elsevier, 1972.

[26] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28:133–168, 1997.

[27] D. Golovin and A. Krause. Adaptive submodularity: A new approach to active learning and stochastic optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2010.

[28] P. Goos and B. Jones. *Optimal Design of Experiments: A Case Study Approach.* Wiley, 2011.

[29] Silviu Guiaşu. *Information Theory with Applications.* McGraw-Hill, New York, 1977.

[30] J. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. 1971.

[31] S. Huang, N. M. Kwok, G. Dissanayake, Q. P. Ha, and G. Fang. Multi-step look-ahead trajectory planning in SLAM: Possibility and necessity. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1091–1096, Piscataway, NJ, April 2005. IEEE.

[32] A. Ihler and D. McAllester. Particle belief propagation. In D. van Dyk and M. Welling, editors, *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 256–263, Clearwater Beach, Florida, 2009. JMLR: W&CP 5.

[33] M. I. Jordan. Graphical models. *Statistical Science*, 19(1):140–155, 2004.

[34] S. Joshi and S. Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57(2):451–462, 2009.

[35] L. Kaelbling, M. Littman, and A. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[36] C. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43:684–691, 1995.

[37] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[38] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, June 2004.

[39] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

[40] A. Krause and C. Guestrin. A note on the budgeted maximization of submodular functions. Technical Report CMU-CALD-05-103, CMU, 2005.

[41] A. Krause and C. Guestrin. Optimal value of information in graphical models. *Journal of Artificial Intelligence Research (JAIR)*, 35:557–591, 2009.

[42] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research (JMLR)*, 9:235–284, 2008.

[43] C. M. Kreucher, A. O. Hero, and K. D. Kastella. An information-based approach to sensor management in large dynamic networks. *Proceedings of the IEEE, Special Issue on Modeling, Identificiation, & Control of Large-Scale Dynamical Systems*, 95(5):978–999, May 2007.

[44] T. O. Kvålseth. The relative useful information measure: Some comments. *Information Sciences*, 56(1-3):35–38, August 1991.

[45] S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, UK, 1996.

[46] J. Le Ny and G. J. Pappas. On trajectory optimization for active sensing in Gaussian process models. In *Proceedings of the IEEE Conference on Decision and Control (CDC)*, pages 6286–6292. IEEE, December 2009.

[47] D. Levine and J. P. How. Sensor selection in high-dimensional Gaussian trees with nuisances. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 26, pages 2211–2219, 2013.

[48] D. Levine and J. P. How. Quantifying nonlocal informativeness in high-dimensional, loopy gaussian graphical models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.

[49] D. S. Levine, B. Luders, and J. P. How. Information-theoretic motion planning for constrained sensor networks. *Journal of Aerospace Information Systems (JAIS)*, 10(10):476–496, 2013.

[50] C. Lieberman. *Goal-oriented inference: Theoretic foundations and application to carbon capture and storage.* PhD thesis, Massachusetts Institute of Technology, June 2013.

[51] D. V. Lindley. On a measure of information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1956.

[52] M. Littman, A. Cassandra, and L. Kaelbling. Learning policies for partially observable environments: scaling up. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 362–370, San Francisco, CA, 1995.

[53] Y. Liu, V. Chandrasekaran, A. Anandkumar, and A. S. Willsky. Feedback message passing for inference in Gaussian graphical models. *IEEE Transactions on Signal Processing*, 60(8):4135–4150, August 2012.

[54] Y. Liu and A. S. Willsky. Learning Gaussian graphical models with observed or latent FVSs. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, 2013.

[55] Y. Liu and A. S. Willsky. Recursive FMP for distributed inference in Gaussian graphical models. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Istanbul, Turkey, July 2013. IEEE.

[56] D. J. Lizotte, O. Madani, and R. Greiner. Budgeted learning of naïve-Bayes classifiers. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2003.

[57] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang. The factor graph approach to model-based signal processing. *Proceedings of the IEEE*, 95(6):1295–1322, June 2007.

[58] K. H. Low, J. M. Dolan, and P. Khosla. Information-theoretic approach to efficient adaptive path planning for mobile robotic environmental sensing. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 2009.

[59] D. M. Malioutov, J. K. Johnson, and A. S. Willsky. Walk-sums and belief propagation in Gaussian graphical models. *Journal of Machine Learning Research (JMLR)*, 7:2031–2064, 2006.

[60] P. S. Maybeck. *Stochastic models, estimation, and control*, volume 1. Academic Press, New York, 1979.

[61] R. McEliece, D. J. C. MacKay, and J. Cheng. Turbo decoding as an instance of pearl's 'belief propagation' algorithm. *IEEE Journal on Selected Areas in Communications*, 16(2):140–152, 1998.

[62] M. McKay, R. Beckman, and W. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245, 1979.

[63] A. Meliou, A. Krause, C. Guestrin, and J. M. Hellerstein. Nonmyopic informative path planning in spatio-temporal models. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, volume 22, pages 602–607, Menlo Park, CA, 2007. AAAI Press.

[64] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14:489–498, 1978.

[65] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37(2):876–904, 2009.

[66] R. Olfati-Saber. Distributed tracking for mobile sensor networks with information-driven mobility. In *Proceedings of the American Control Conference (ACC)*, pages 4606–4612, July 2007.

[67] J. Oxley. *Matroid Theory*. Oxford University Press, Oxford, 1992.

[68] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufman, San Mateo, CA, 1988.

[69] K. H. Plarre and P. R. Kumar. Extended message passing algorithm for inference in loopy Gaussian graphical models. *Ad Hoc Networks*, 2:153–169, 2004.

[70] R. Platt, Jr., R. Tedrake, L. Kaelbling, and T. Lozano-Perez. Belief space planning assuming maximum likelihood observations. In *Proceedings of Robotics: Science and Systems*, 2010.

[71] A. C. Pocock. *Feature Selection via Joint Likelihood.* PhD thesis, University of Manchester, 2012.

[72] S. S. Ponda, R. M. Kolacinski, and Emilio Frazzoli. Trajectory optimization for target localization using small unmanned aerial vehicles. In *AIAA Guidance, Navigation, and Control Conference (GNC)*, Chicago, IL, August 2009.

[73] S. Prentice and N. Roy. The belief roadmap: Efficient planning in belief space by factoring the covariance. *International Journal of Robotics Research (IJRR)*, 28(11-12):1448–1465, 2009.

[74] N. Ramakrishnan, C. Bailey-Kellogg, S. Tadepalli, and V. Pandey. Gaussian processes for active data mining of spatial aggregates. In *Proceedings of the SIAM International Conference on Data Mining*, 2005.

[75] B. Ristic, M. Morelande, and A. Gunatilaka. Information driven search for point sources of gamma radiation. *Signal Processing*, 90(4):1225–1239, April 2010.

[76] N. Roy, G. Gordon, and S. Thrun. Finding approximate POMDP solutions through belief compression. *Journal of Artificial Intelligence Research (JAIR)*, 23:1–40, 2005.

[77] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 441–448, 2001.

[78] B. Settles. Active learning literature survey. Computer Sciences 1648, University of Wisconsin-Madison, 2009.

[79] R. Sim and N. Roy. Global A-optimal robot exploration in SLAM. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 661–666, Piscataway, NJ, 2005. IEEE.

[80] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M.I. Jordan, and S.S. Sastry. Kalman filtering with intermittent observations. *IEEE Transactions on Automatic Control*, 49(9):1453–1464, September 2004.

[81] L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15, pages 707–715, Ft. Lauderdale, FL, May 2011.

[82] T. P. Speed and H. T. Kiiveri. Gaussian Markov distributions over finite graphs. *Annals of Statistics*, 14(1):138–150, March 1986.

[83] E. B. Sudderth. Embedded trees: Estimation of Gaussian processes on graphs with cycles. Master's thesis, Massachusetts Institute of Technology, February 2002.

[84] E. B. Sudderth, A. T. Ihler, M. Isard, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. *Communications of the ACM*, 53:95–103, October 2010.

[85] E. B. Sudderth, M. J Wainwright, and A. S. Willsky. Embedded trees: Estimation of Gaussian processes on graphs with cycles. *IEEE Transactions on Signal Processing*, 52(11):3136–3150, November 2004.

[86] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. Intelligent Robotics and Autonomous Agents. MIT Press, Cambridge, MA, 2005.

[87] S. Tong and D. Koller. Active learning for parameter estimation in Bayesian networks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 13, 2000.

[88] S. Tong and D. Koller. Active learning for structure in Bayesian networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

[89] J. van den Berg, S. Patil, and R. Alterovitz. Motion planning under uncertainty using iterative local optimaization in belief space. *International Journal of Robotics Research*, 31(11):1263–1278, 2012.

[90] M. J Wainwright, E. B. Sudderth, and A. S. Willsky. Tree-based modeling and estimation of Gaussian processes on graphs with cycles. In T. K. Leen, T. G.

Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 13. MIT Press, November 2000.

[91] Y. Weiss and W. T. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*, 13(10):2173–2200, 2001.

[92] M. Welling and Y. W. Teh. Linear response algorithms for approximate inference in graphical models. *Neural Computation*, 16(1):197–221, 2004.

[93] J. L. Williams. *Information Theoretic Sensor Management.* PhD thesis, Massachusetts Institute of Technology, 2007.

[94] J. L. Williams, J. W. Fisher III, and A. S. Willsky. Performance guarantees for information theoretic active inference. In M. Meila and X. Shen, editors, *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 616–623, 2007.

[95] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. Technical Report TR2001-16, Mitsubishi Research Laboratories, May 2001.

[96] D. M. Young. *Iterative Solution of Large Linear Systems.* Academic, New York, 1971.

The greater the scientist, the more he is impressed with his ignorance of reality, and the more he realizes that his laws and labels, descriptions and definitions, are the products of his own thought. They help him to use the world for purposes of his own devising rather than to understand and explain it.

The more he analyzes the universe into infinitesimals, the more things he finds to classify, and the more he percieves the relativity of all classification. What he does not know seems to increase in geometric progression to what he knows. Steadily, he approaches the point where what is unknown is not a mere blank space in a web of words, but a window in the mind, a window whose name is not ignorance but wonder.

Alan Watts, *The Wisdom of Insecurity*