Constrained Information Exchange Message Passing Algorithm in Probabilistic Graphical Models

By

Mohamed Ibrahim AlHajri

B.S., Khalifa University (2015)
S.M., Massachusetts Institute of Technology (2018)

Submitted to the Department of Electrical Engineering and Computer Science in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2023

Authored by: Mohamed Ibrahim AlHajri
Department of Electrical Engineering and Computer Science
June 13, 2023

Certified by: Lizhong Zheng
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by: Raed M. Shubair
Professor of Electrical and Computer Engineering, New York University Abu Dhabi
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Constrained Information Exchange Message Passing Algorithm in Probabilistic Graphical Models

by

Mohamed Ibrahim AlHajri

Submitted to the Department of Electrical Engineering and Computer Science
on June 13, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Probabilistic graphical models combine probability and graph theory into a powerful multivariate statistical modeling approach. While there is an extraordinary range of types of graphical models in the broader literature, we focus on Hidden Markov Model (HMM) and Linear Dynamical System (LDS). These two models are ubiquitous in applications including Kalman filtering, and the decoding of LDPC codes which is what all cellular wireless systems run on these days. Message passing algorithms exploit the independence and factorization structure within these graphical models to develop analytically tractable, computationally efficient, and exact inference algorithms. Posterior inference using message-passing algorithm depends heavily on the information exchange between the nodes of the graph and having unconstrained sized messages, results in the exact inference of the posterior distribution. Despite its usefulness and ubiquitous applications, the unconstrained information exchange message-passing algorithm has numerous shortcomings, including prohibitive computational complexity, and unaffordable communication complexity. These shortcomings can limit the effectiveness and practicality of the algorithm in the era of big data.

This thesis presents a comprehensive analysis of a novel constrained information exchange message-passing algorithm. A cornerstone of this algorithm is the modified sum product algorithm, an innovative approach that identifies and prioritizes critical information fragments for particular inference tasks.

The thesis further delves into the algorithm's utility in various posterior inference scenarios, encompassing index-specific and index-free posterior inferences. While the former scrutinizes either singular or multiple posterior inferences at designated indexes, the latter ensures that all nodes use identical compression matrices, catering to both single and multi-step posterior inferences.

The algorithm elucidates the balance between algorithmic performance and resource utilization in data-driven applications. By reducing computational and communication complexities, it proposes an efficient solution for high-dimensional data handling, particularly when resources are constrained. In essence, the proposed algorithm enhances the scalability, practicality, and efficiency of probabilistic graphical models, propelling advancements in data-driven research and decision-making across a multitude of domains.

Thesis Supervisor: Lizhong Zheng
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Raed Shubair
Title: Professor of Electrical Engineering

Acknowledgments

Expressing my love to my parents will never be enough. Their infinite love, support, and sacrifice has always been behind every success I achieved over the years. Their enlightening wisdom and continuous encouragement were essential for the completion of my PhD degree. Their love will always shine in my heart and their faces will always brighten my eyes. I would like to express my sincere gratitude to my brothers and sisters (Ahmed, Hamad, Mariam, Noura, Lamia). Their love, trust, and encouragement have always been a source of energy, confidence, and determination. I am grateful for having such a great family, my heart will always be with you.

I would like to express my sincere gratitude and deepest appreciation to my thesis supervisor, Prof. Lizhong Zheng, for giving me the opportunity to work with him as a member of the Claude E. Shannon Communication and Network Research Group. I am very grateful for his continuous support, guidance, and kind advice. Prof. Zheng was always there when I had any doubt and has constantly supported me throughout my journey. Working with him has helped me gain a lot of insight when approaching sophisticated problems. Prof. Zheng always emphasized the importance of keeping my focus on the concept rather than the details. I have learned a lot from his clarity, vision, and genuine approach. He will always be a source of admiration.

I am heartily thankful and grateful to my thesis supervisior, Prof. Raed Shubair for his continuous encouragement and unlimited support, which started while I was an undergraduate student in UAE and continued during my postgraduate studies at MIT. Prof. Shubair is an amazing mentor and an exceptional motivator. Working closely with Prof. Shubair has helped me sharpen my critical thinking and gain confidence in tackling challenging problems. His timeless help and relentless guidance have always been essential for my progress and success. I enjoyed many lovely memories with Prof. Shubair. Prof. Shubair's dedication and persistence will always remain a source of inspiration.

I am immensely grateful to Prof. Vincent Chan for his invaluable contribution as a member of my PhD thesis committee. His valuable help and generous support greatly impacted my life in immeasurable ways.

I am extremely indebted to Prof. Muriel Medard for her valuable help and kind support. Her warm welcome when I first joined MIT along with her honest advice made me feel more comfortable at the Research Laboratory of Electronics (RLE).

I am also thankful to Prof. Reyad El-Khazali and Prof. Luis Weruaga who had an instrumental role during my undergraduate days and have always encouraged me to pursue my graduate studies.

I would like to thank all my friends Fahad Alhasoun, Ali Al Jabri, Abdulla Alomar, Ali Aljefri, Mohamad Alrished, Abdulatif aljarbou, Abdulaziz Alhassan, Abdullah Alfaqih, Mohammed Alsobay, Abdulrahman Alotaibi, Abdulla Alhajri, Saeed Al Jaberi, and many more that have been a source of wisdom and entertainment.

I would like to thank my former and current mates in the Claude E. Shannon Communication and Network Research Group, Mina Karzand, Fabian Kozynski, Anuran Makur, David Qiu, Jiejun Jiu, Erixhen Sula, Xiangxiang Xu, Melichan Erol, and Eric Wang for the interesting discussions and insightful advices. I will cherish moments of getting together to celebrate beautiful occasions. I wish you all an amazing future!

A heartfelt thanks to the UAE Presidential Court and the Scholarship Office for sponsoring me through the "Distinguished Student Scholarship of His Highness the President of the United Arab Emirates".

I would like to thank the staff in the Research Laboratory of Electronics (RLE) and at the Department of Electrical Engineering and Computer Science (EECS) for providing such a delightful environment and a lifetime experience.

To the loving memory of

my grandparents,

for their love, support, and values.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Navigating Uncertainty

To achieve their goals, intelligent agents, whether natural or artificial, must choose from numerous possible courses of action. They must make decisions based on information obtained from their environment, prior knowledge, and objectives. In many cases, this information and knowledge may be incomplete or unreliable, necessitating decision-making under uncertainty. For example, in an emergency, a medical doctor must act quickly despite having limited information about the patient's condition; an autonomous vehicle detecting a potential obstacle must decide whether to turn or stop without being certain of the obstacle's distance, size, and speed.

A goal of artificial intelligence is to create systems capable of reasoning and making decisions under uncertain conditions. Early intelligent systems faced challenges in handling uncertainty, as traditional paradigms were ill-equipped to manage it.

Classical logic-based early artificial intelligence systems represented knowledge as sets of logical clauses or rules. These systems exhibited two crucial properties *modularity* and *monotonicity* which simplified knowledge acquisition and inference [1].

A system is considered *modular* if each piece of knowledge can independently lead to conclusions. In other words, if the premises of any logical clause or rule are true, its conclusion can be asserted without considering other elements in the knowledge base. A system exhibits *monotonicity* if its knowledge consistently increases monotonically, meaning any deduced fact or conclusion is maintained even when new facts become known to the system.

However, when uncertainty is involved, these properties generally do not hold true. The absence of these properties makes reasoning more complex for a system. In principle, the system must consider all available knowledge and facts when drawing a conclusion and be prepared to revise its conclusions as new data emerges. To address this challenge, modern artificial intelligence systems employ probabilistic reasoning, machine learning algorithms, and other approaches to better manage uncertainty, adapt to changing conditions, and make more informed decisions.

1.2 Probability Theory

Probability theory offers a well-established foundation for managing uncertainty, making it a natural choice for reasoning under uncertain conditions. However, naively applying probability to complex problems quickly results in overwhelming computational complexity.

Consider a collection of random variables $\mathbf{x} = (\mathsf{x}_1, ..., \mathsf{x}_N)$ and let the observations about them be represented by random variables $\mathbf{y} = (\mathsf{y}_1, ..., \mathsf{y}_N)$. Let each of these random variables x_i , $1 \leq i \leq N$, take on a value in $\mathcal{X}$ (e.g., $\mathcal{X} = \{0, 1\}$ or $\mathbb{R}$) and each observation variable y_i , $1 \leq i \leq N$ take on a value in $\mathcal{Y}$. Given observation $\mathbf{y}$, the goal is to say something meaningful about possible realizations of $\mathbf{x}$ for $\mathbf{x} \in \mathcal{X}^N$. One can consider there to be effectively two primary computation problems of interest, as we now describe.

Calculating (Posterior) Beliefs and Marginalization. Here the goal is to perform computations of the form

$$p_{\mathbf{x}|\mathbf{y}}(x_1^n|\hat{y}_1^n) = \frac{p_{\mathbf{x},\mathbf{y}}(x_1^n, \hat{y}_1^n)}{p_{\mathbf{y}}(\hat{y}_1^n)} = \frac{p_{\mathbf{x},\mathbf{y}}(x_1^n, \hat{y}_1^n)}{\sum_{\mathsf{x}_1,\cdots,\mathsf{x}_n \in \mathcal{X}} p_{\mathbf{x},\mathbf{y}}(x_1^n, \hat{y}_1^n)} \tag{1.1}$$

The denominator (i.e., normalization) is referred to as the partition function.

Evidently, calculating the partition function corresponds to executing marginalization, which can be excessively costly for larger alphabets, as commonly seen in high-dimensional situations, unless there is a specific structure that can be utilized.

Assuming the objective is to marginalize M components out of the N elements of $\mathbf{x}$, producing a marginal of $(N-M)$ dimensions, it is observed that each summation in this process requires about $|\mathcal{X}|^M$ computations. Moreover, this has to be executed for each of the $|\mathcal{X}|^{(N-M)}$ potential values of the remaining variables. The total complexity hence amounts to approximately $|\mathcal{X}|^N$, which is exponential in the number of variables N. This typically becomes unmanageable for even moderately large values of N.

It's worth noting that this outcome is not unexpected. In the absence of a structure that can be leveraged, the joint distribution of a set of N variables, each over the alphabet $|\mathcal{X}|$, would necessitate a table of $|\mathcal{X}|^N$. Hence, even merely accessing all the values of the joint distribution would involve exponential complexity.

Calculating Most Probable Configurations (MPC). The goal is to perform the following computation:

$$\hat{\mathbf{x}} \in \arg\max_{\mathbf{x}\in\mathcal{X}^N} p_{\mathbf{x}|\mathbf{y}}(x_1^n|\hat{y}_1^n) \tag{1.2}$$

$$\hat{\mathbf{x}} \in \arg\max_{\mathbf{x}\in\mathcal{X}^N} p_{\mathbf{x}|\mathbf{y}}(x_1^n|\hat{y}_1^n) = \arg\max_{\mathbf{x}\in\mathcal{X}^N} \frac{p_{\mathbf{x},\mathbf{y}}(x_1^n,\hat{y}_1^n)}{p_{\mathbf{y}}(\hat{y}_1^n)} = \arg\max_{\mathbf{x}\in\mathcal{X}^N} p_{\mathbf{x},\mathbf{y}}(x_1^n,\hat{y}_1^n) \tag{1.3}$$

Without any additional structure, the above optimization problem obviously requires searching over all $\mathcal{X}^N$ entries in the table representing the joint distribution $p_{\mathbf{x}|\mathbf{y}}(.|\hat{y}_1^n)$, so has complexity that is exponential in N.

To overcome the computational complexity associated with probability-based reasoning, probabilistic graphical models have been introduced to exploit the structure in joint distributions, reducing complexity. These models enable more efficient representation and manipulation of probability distributions by utilizing graphs to illustrate the conditional dependencies between variables, which simplifies calculations and makes it possible to reason under uncertainty in a computationally feasible manner.

1.3 Probabilistic Graphical Models

Probabilistic Graphical Models (PGM) are robust statistical modeling techniques that blend the principles of probability and graph theory. They offer a strategic framework for handling uncertainty, grounded in probability theory, and enable efficient computational processes [2]. The core concept involves recognizing and focusing on the independence relations pertinent to a specific problem, integrating these into the probabilistic model to curtail complexity, and thus minimizing memory needs and computational time.

A PGM serves as a concise depiction of a joint probability distribution, from which we can derive both marginal and conditional probabilities. Its power lies in representing the dependencies and independencies among a group of variables using graphs. In these graphs, variables exhibiting direct dependence are linked, while the independence relations are subtly embedded within this dependency graph.

Essentially, PGMs efficiently capture and represent complex probabilistic relationships, providing a simplified yet comprehensive view of the problem at hand. This clarity and reduction in complexity make PGMs an extremely valuable tool for various fields requiring probabilistic inference and decision-making under uncertainty.

To illustrate the dramatic impact structure can have, consider the following example. Specifically, suppose now that the N random variables of interest are mutually independent, so that

$$p_{\mathsf{x}_1,\cdots,\mathsf{x}_N}(x_1,\cdots,x_N) = p_{\mathsf{x}_1}(x_1)p_{\mathsf{x}_2}(x_2)\cdots p_{\mathsf{x}_N}(x_N) \tag{1.4}$$

When computing posterior beliefs, the marginalization can be performed individually for each variable. Each of these computations has a complexity of $|\mathcal{X}|$, leading to a total complexity on the order of $N|\mathcal{X}|$, which is linear in N. In a similar vein, the most probable configurations can be determined by independently identifying the assignment of each variable that maximizes its own probability. Given that there are N variables, the overall computational complexity is also linear in N, i.e., of order $N|\mathcal{X}|$.

Clearly, the presence of an independence structure can significantly reduce the complexity of inference tasks. More broadly, more intricate forms of independence and related structures in distributions can be leveraged to similarly reduce the complexity of these foundational inference tasks, often to a surprisingly high degree.

1.4 Limitations of Probabilistic Graphical Models

Message-passing algorithm depends heavily on the information exchange between the nodes of the graph and having unconstrained sized messages, results in the exact inference of the posterior distribution. Despite its usefulness and ubiquitous applications, the unconstrained information exchange message-passing algorithm has numerous shortcomings, including (1) prohibitive computational complexity, and (2) unaffordable communication complexity. These shortcomings can limit the effectiveness and practicality of the algorithm in modern applications:

- **High-dimensional data**: The random variable involved is no longer the output of an individual sensor or a sample of a waveform but more often an image or a sentence in natural language represented in a vector space with multiple thousands of dimensions.

- **Limited Communication Bandwidth**: The sensors used to capture the data in a wide range of applications such as smart cities [3], environment monitoring [4], security and surveillance [5], industrial monitoring [6, 7], and many more have limited communication capability.

In these problems, on top of the savings by the graphical models, we also need additional savings by using the dependence between a single pair of random variables. In this thesis, we address these new challenges with very high-dimensional data or limited communication bandwidth by introducing and analyzing the **constrained information exchange message passing algorithm**.

Constraints are placed on the messages one can pass along in the algorithm and deliberately sets it to be less than what is needed to carry sufficient statistics. This formulation forces us to look for "insufficient statistics," which we expect to be the key new concept in the era of big data as it will result in lower computational and communication complexity. We will have to learn how to compromise by dropping the less important parts of our information, which requires a better understanding of quantitative metrics of how important fragments of information are and how much they can be helpful with specific inference tasks.

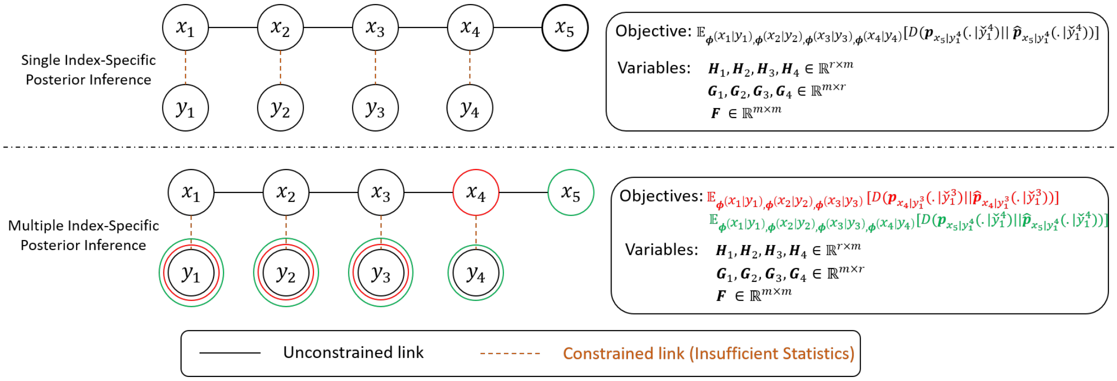
The restrictive nature of the constraints, which limit the exchange of sufficient statistics, inherently complicates the process of posterior inference. This complexity arises from the unique dynamics of the process, where a distinct insufficient statistic is forwarded as a message, depending on the specific scenario of the posterior inference. Consequently, this intricate issue not only amplifies the complexity of the problem but also necessitates the crafting of diverse messages tailored to each unique posterior inference scenario. In order to address these challenges, it becomes crucial to develop and rigorously analyze the algorithm under a broad range of scenarios. By doing so, one can ensure the algorithm's robustness and adaptability, allowing it to effectively handle a multitude of posterior inference conditions, regardless of the constraints imposed on the information exchange.

1.5 Outline of the Thesis

The purpose of this thesis is to propose and analyze constrained information exchange message passing algorithm and this is where insufficient statistic messages are used in the algorithm. This formulation will be of crucial importance to deal with the new challenges introduced with the high-dimensional data.

Chapter 2 gives an overview of the hidden markov model and then discusses the message passing algorithm for posterior inference. Then, a modified sum product algorithm will be introduced in terms of the information vector and Divergence Transfer Matrix (DTM). This modification allows for the determination of which fragments of information are most important for specific inference tasks. By identifying these important pieces of information, informed decisions can be made about which parts of the data to prioritize and which parts can be safely dropped or simplified.

Chapter 3 introduces the index-specific posterior inference using constrained information exchange message passing algorithm. This sub-category of posterior inference provides a critical examination of specific scenarios where the interest lies either in a singular posterior inference at a specific index (referred to as single index-specific posterior inference) or in multiple posterior inferences at specific indexes (referred to as multiple index-specific posterior inference) as shown in Fig. 1-1.



Single Index-Specific Posterior Inference
x_1
x_2
x_3
x_4
x_5
y_1
y_2
y_3
y_4
Objective: $\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2),\boldsymbol{\phi}(x_3|y_3),\boldsymbol{\phi}(x_4|y_4)}[D(\boldsymbol{p}_{x_5|y_1^4}(.|\check{y}_1^4)||\,\hat{\boldsymbol{p}}_{x_5|y_1^4}(.|\check{y}_1^4))]$
Variables: $\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3,\boldsymbol{H}_4 \in \mathbb{R}^{r\times m}$
$\boldsymbol{G}_1,\boldsymbol{G}_2,\boldsymbol{G}_3,\boldsymbol{G}_4 \in \mathbb{R}^{m\times r}$
$\boldsymbol{F} \in \mathbb{R}^{m\times m}$
Multiple Index-Specific Posterior Inference
x_1
x_2
x_3
x_4
x_5
y_1
y_2
y_3
y_4
Objectives: $\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2),\boldsymbol{\phi}(x_3|y_3)}[D(\boldsymbol{p}_{x_4|y_1^3}(.|\check{y}_1^3)||\hat{\boldsymbol{p}}_{x_4|y_1^3}(.|\check{y}_1^3))]$
$\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2),\boldsymbol{\phi}(x_3|y_3),\boldsymbol{\phi}(x_4|y_4)}[D(\boldsymbol{p}_{x_5|y_1^4}(.|\check{y}_1^4)||\hat{\boldsymbol{p}}_{x_5|y_1^4}(.|\check{y}_1^4))]$
Variables: $\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3,\boldsymbol{H}_4 \in \mathbb{R}^{r\times m}$
$\boldsymbol{G}_1,\boldsymbol{G}_2,\boldsymbol{G}_3,\boldsymbol{G}_4 \in \mathbb{R}^{m\times r}$
$\boldsymbol{F} \in \mathbb{R}^{m\times m}$
Unconstrained link
Constrained link (Insufficient Statistics)

Figure 1-1: Index-Specific Posterior Inference Using Constrained Information Exchange Message-Passing Algorithm

Chapter 4 introduces the index-free posterior inference using constrained information exchange message passing algorithm. Unlike the index-specific approach, where nodes may have different compression matrices, the index-free approach ensures that all nodes employ identical compression matrices. This sub-category of posterior inference provides a critical examination of specific scenarios where the interest lies either in a single step posterior inference (referred to as single step index-free posterior inference) or in multi-step posterior inferences (referred to as multi-step index-free posterior inference) as shown in Fig. 1-2.

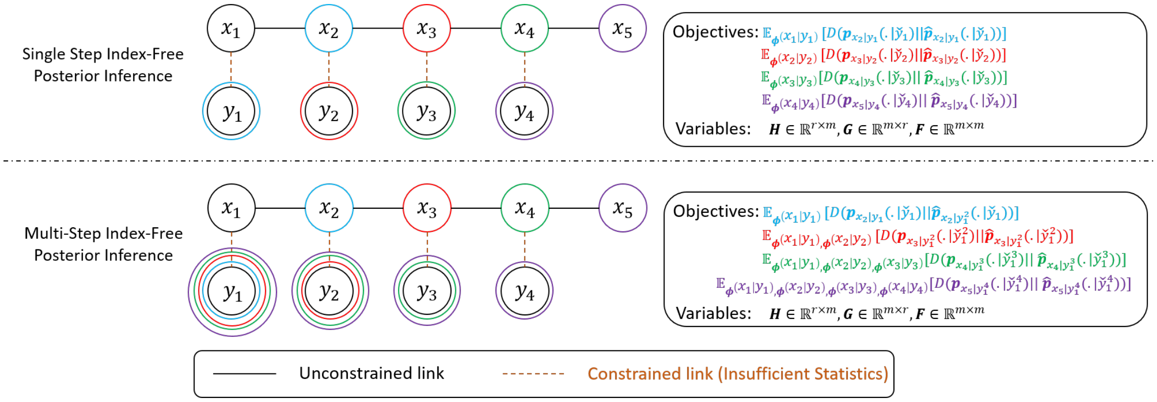

Single Step Index-Free Posterior Inference
x_1
x_2
x_3
x_4
x_5
y_1
y_2
y_3
y_4
Objectives: $\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1)}[D(\boldsymbol{p}_{x_2|y_1}(.|\check{y}_1)||\widehat{\boldsymbol{p}}_{x_2|y_1}(.|\check{y}_1))]$
$\mathbb{E}_{\boldsymbol{\phi}(x_2|y_2)}[D(\boldsymbol{p}_{x_3|y_2}(.|\check{y}_2)||\widehat{\boldsymbol{p}}_{x_3|y_2}(.|\check{y}_2))]$
$\mathbb{E}_{\boldsymbol{\phi}(x_3|y_3)}[D(\boldsymbol{p}_{x_4|y_3}(.|\check{y}_3)||\widehat{\boldsymbol{p}}_{x_4|y_3}(.|\check{y}_3))]$
$\mathbb{E}_{\boldsymbol{\phi}(x_4|y_4)}[D(\boldsymbol{p}_{x_5|y_4}(.|\check{y}_4)||\widehat{\boldsymbol{p}}_{x_5|y_4}(.|\check{y}_4))]$
Variables: $\boldsymbol{H} \in \mathbb{R}^{r\times m}, \boldsymbol{G} \in \mathbb{R}^{m\times r}, \boldsymbol{F} \in \mathbb{R}^{m\times m}$
Multi-Step Index-Free Posterior Inference
x_1
x_2
x_3
x_4
x_5
y_1
y_2
y_3
y_4
Objectives: $\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1)}[D(\boldsymbol{p}_{x_2|y_1}(.|\check{y}_1)||\widehat{\boldsymbol{p}}_{x_2|y_1^2}(.|\check{y}_1))]$
$\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2)}[D(\boldsymbol{p}_{x_3|y_1^2}(.|\check{y}_1^2)||\widehat{\boldsymbol{p}}_{x_3|y_1^2}(.|\check{y}_1^2))]$
$\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2),\boldsymbol{\phi}(x_3|y_3)}[D(\boldsymbol{p}_{x_4|y_1^3}(.|\check{y}_1^3)||\widehat{\boldsymbol{p}}_{x_4|y_1^3}(.|\check{y}_1^3))]$
$\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2),\boldsymbol{\phi}(x_3|y_3),\boldsymbol{\phi}(x_4|y_4)}[D(\boldsymbol{p}_{x_5|y_1^4}(.|\check{y}_1^4)||\widehat{\boldsymbol{p}}_{x_5|y_1^4}(.|\check{y}_1^4))]$
Variables: $\boldsymbol{H} \in \mathbb{R}^{r\times m}, \boldsymbol{G} \in \mathbb{R}^{m\times r}, \boldsymbol{F} \in \mathbb{R}^{m\times m}$
Unconstrained link
Constrained link (Insufficient Statistics)

Figure 1-2: Index-Free Posterior Inference Using Constrained Information Exchange Message-Passing Algorithm

Chapter 5 concludes the presented work and outlines future directions.

Chapter 2

Hidden Markov Model

2.1 Overview of HMM

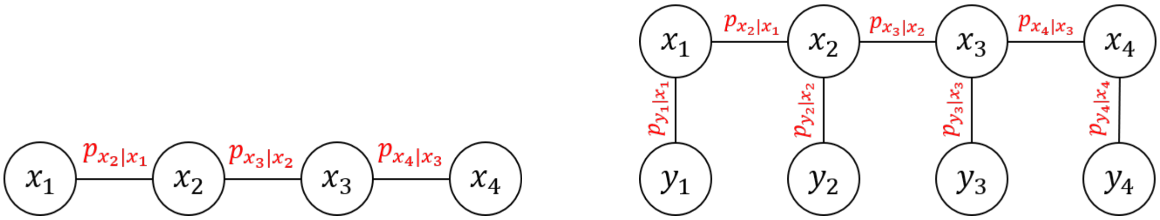
Markov chains as shown in Fig. 2-1a, are a type of mathematical model used to describe systems that transition between different states over time [8]. The key feature of a Markov chain is that the probability of transitioning from one state to another only depends on the current state of the system, and not on any previous states. This property is known as the Markov property, and it allows us to model complex systems in a simple and efficient way [8].

In a Markov chain, the transition probabilities between states are typically represented as a transition matrix. The transition matrix describes the probability of transitioning from one state to another, given the current state of the system. These probabilities are often estimated from data, or from expert knowledge about the system being modeled.

One interesting application of Markov chains is in the field of latent variable models, which are models that involve variables that are not directly observable. Hidden Markov models (HMMs) as shown in Fig. 2-1b are one example of a latent variable model that is widely used in many different fields.

In an HMM, the system is modeled as a Markov chain, where the hidden state of the system is represented by a set of states that are not directly observable [9]. Instead, the system generates observations that are dependent on the current hidden state, according to a set of emission probabilities. These emission probabilities describe the probability of observing a particular observation, given the current hidden state of the system. The transition probability matrix will be defined as $\mathbf{P}_{\mathsf{x}_{i+1}|\mathsf{x}_i}$ and the emission probability matrix will be defined as $\mathbf{P}_{\mathsf{y}_i|\mathsf{x}_i}$.

HMMs are a type of probabilistic model, meaning that they represent the uncertainty inherent in many real-world systems. By modeling the system in this way, HMMs can be used to infer the underlying state of the system based on observed data, and to predict future states.



(a) Markov Model

(b) Hidden Markov Model

2.1.1 Applications

2.1.1.1 Automatic Speech Recognition

Hidden Markov Models (HMMs) are extensively utilized in automatic speech recognition, a technology integral to numerous applications like digital assistants on smartphones, transcription services, and assistive technologies for the hearing impaired [10]. The process of automatic speech recognition involves breaking down the audio input into time intervals, typically within a range of 10 to 30 milliseconds. Each of these segments is analyzed to capture its unique acoustic features. HMMs excel in this task due to their ability to model the temporal structure of speech sounds, thereby accommodating for variations in pronunciation, intonation, rhythm, and other acoustic nuances that characterize individual speech patterns [11].

In an automatic speech recognition system powered by HMMs, the audio waveform serves as the observed data, while the hidden states correspond to the phonemes, which are the smallest units of sound that make up the words in the spoken language [12]. The transition probabilities between these hidden states encapsulate the likelihood of one phoneme following another in the language, thereby reflecting linguistic rules and common patterns of the language. On the other hand, the emission probabilities depict the probability of observing a specific acoustic feature given a particular phoneme [13].

This application of HMMs in automatic speech recognition underpins the transformation of human-computer interaction, providing a more intuitive and accessible way for users to engage with technology [14].

2.1.1.2 Natural Language Processing

Hidden Markov Models (HMMs) play an instrumental role in Natural Language Processing (NLP), a subfield of artificial intelligence concerned with the interaction between computers and human language. Tasks like part-of-speech tagging, and named entity recognition are areas where HMMs have been successfully deployed [15].

Part-of-Speech (POS) tagging is the process of assigning a grammatical category, such as noun, verb, adjective, etc., to each word in a sentence. HMMs are often used in POS tagging, where the hidden states represent the grammatical categories, and the observed states represent the words. The transition probabilities between the hidden states capture the grammatical rules of the language, while the emission probabilities capture the likelihood of a word belonging to a particular grammatical category [16–18].

Named Entity Recognition (NER) is another NLP task that involves identifying and classifying named entities in a text into predefined categories such as person names, organizations, locations, etc. HMMs have been used for NER by modeling the dependencies between neighboring words and their corresponding entity labels. In this case, the hidden states represent the entity labels, and the observed states represent the words [19].

2.1.1.3 Genetics and Bioinformatics

Hidden Markov Models (HMMs) have emerged as a crucial tool in genetic sequence analysis and bioinformatics, particularly in protein family characterization, gene prediction, and sequence alignment [20]. This relevance of HMMs is largely due to the inherently sequential nature of genetic data, which bears a striking resemblance to the temporal structure modeled by HMMs [21].

In the realm of bioinformatics, the observed data corresponds to the genetic sequence, typically represented as a string of nucleotides or amino acids. The hidden states, on the other hand, represent biological structures or functional regions, such as exons, introns, intergenic regions, or different structural motifs in proteins [22]. The transition probabilities in these HMMs signify the likelihood of moving from one biological structure to another in the genome or protein, thereby embodying the underlying biological processes and mechanisms [23].

Emission probabilities in this context denote the likelihood of observing a particular nucleotide or amino acid given a specific biological structure or functional region [24]. Hence, these probabilities encompass the inherent variability within each biological structure, enabling HMMs to model the biological complexity and variation in genetic sequences effectively.

HMMs have been instrumental in transforming our understanding of genomics and proteomics, fostering the development of algorithms and databases that enhance the prediction and classification of genes and proteins [25, 26]. The application of HMMs in bioinformatics also fosters the identification of potential therapeutic targets and the development of personalized medicine, providing a more detailed and comprehensive understanding of genetic mechanisms underlying disease [27].

2.1.1.4 Wireless Communication

Hidden Markov Models (HMMs) have found significant application in wireless communication systems, playing a crucial role in tasks like channel equalization, and signal decoding [28]. These applications leverage the temporal modelling capabilities of HMMs, which excellently suit the dynamic and sequential nature of wireless signal transmission.

In the context of wireless communication, the observed data typically corresponds to the received signal, which is generally a noisy version of the transmitted signal.

The hidden states can represent the transmitted data symbols or the different states of the communication channel [29]. The transition probabilities between the hidden states capture the dynamics of the channel or the correlation structure of the data symbols. On the other hand, the emission probabilities encapsulate the statistical relationship between the transmitted symbols and the received signals [30].

Channel equalization is one significant application of HMMs in wireless communication, where the goal is to mitigate the adverse effects of channel impairments such as multipath fading and inter-symbol interference [31]. By modeling the communication channel as a hidden Markov process, HMMs enable effective equalization strategies that outperform traditional linear equalization techniques [32].

In signal decoding, HMMs are used to estimate the most likely sequence of transmitted data symbols given the received signals, helping to recover the original data in the presence of noise and channel errors [33]. This has been pivotal in enhancing the reliability and robustness of wireless communication systems.

Through these applications, HMMs have proven to be instrumental in advancing wireless communication technology, fostering improvements in data transmission quality, speed, and reliability.

2.1.1.5 Wireless Sensor Networks

Hidden Markov Models (HMMs) are instrumental in optimizing the performance of wireless sensor networks (WSNs), a technology critical to various applications such as environmental monitoring, healthcare, smart homes, and industrial automation [34]. These networks consist of spatially distributed sensors that cooperatively monitor physical or environmental conditions and communicate their data through the network to a main location. The role of HMMs in this context includes but not limited to tasks such as sensor data fusion, anomaly detection, prediction and network state estimation [35–37].

In WSNs employing HMMs, the observed data can correspond to sensor readings while the hidden states can represent the underlying physical phenomena or the status of the network nodes. Transition probabilities between these hidden states capture the temporal dynamics of the environment or the state changes in the network, and the emission probabilities capture the likelihood of obtaining a particular sensor reading given the underlying state [38].

Moreover, HMMs are utilized in WSNs for detecting anomalies or unusual events. By learning the normal patterns of sensor readings and their transitions, HMMs can identify deviations from these patterns as potential anomalies, thereby contributing to the robustness and security of these networks [39].

2.2 Sum-Product Algorithm

The sum-product algorithm is a message passing algorithm that is used to compute optimally the posterior probabilities of the hidden states in a hidden Markov model (HMM). The sum-product algorithm works by computing and forwarding sufficient statistics at each step as messages. These sufficient statistics are the transition and emission probabilities, given the observations and the current estimates of the model parameters.

The sum-product algorithm is optimal in the sense that it computes the exact posterior probabilities of the hidden states, given the available observations and the model parameters.

For a hidden markov model, we define the node potential and the edge potential as

$$\eta_{\mathsf{x}_i}(x_i) \triangleq p_{\mathsf{x_i}}(x_i) \tag{2.1}$$

$$\eta_{\mathsf{y}_i}(y_i) \triangleq p_{\mathsf{y_i}}(y_i) \tag{2.2}$$

$$\xi_{\mathsf{x}_i,\mathsf{x}_j}(x_i, x_j) \triangleq \frac{p_{\mathsf{x_i},\mathsf{x_j}}(x_i, x_j)}{p_{\mathsf{x_i}}(x_i)p_{\mathsf{x_j}}(x_j)} \tag{2.3}$$

$$\xi_{\mathsf{x}_i,\mathsf{y}_i}(x_i, y_i) \triangleq \frac{p_{\mathsf{x_i},\mathsf{y_i}}(x_i, y_i)}{p_{\mathsf{x_i}}(x_i)p_{\mathsf{y_i}}(y_i)} \tag{2.4}$$

This setup works for HMM, in the sense that the normal definition of the joint distribution over the HMM is consistent with the product of all potentials.

Suppose with observation of $\mathsf{y_1, \ldots, y_{n-1}} = \check{y}_1, \ldots, \check{y}_{n-1}$, and we would like to predict the value of x_n. The forward message from the observation node to the hidden node $m_{\mathsf{y}_i \to \mathsf{x}_i}(x_i)$ is:

$$m_{\mathsf{y}_i \to \mathsf{x}_i}(x_i) = \sum_{y_i} \eta_{\mathsf{y}_i}(y_i) \xi_{\mathsf{x}_i, \mathsf{y}_i}(x_i, y_i) \mathbf{1}(y_i = \check{y}_i) = p_{\mathsf{y}_i|\mathsf{x}_i}(\check{y}_i | x_i) \tag{2.5}$$

The first forward message between the hidden nodes $m_{\mathsf{x}_1 \to \mathsf{x}_2}(x_2)$ depends on $m_{\mathsf{y}_1 \to \mathsf{x}_1}(x_1)$, $\eta_{\mathsf{x}_1}(x_1)$, and $\xi_{\mathsf{x}_1, \mathsf{x}_2}(x_1, x_2)$.

$$\begin{aligned} m_{\mathsf{x}_1 \to \mathsf{x}_2}(x_2) &= \sum_{x_1} \eta_{\mathsf{x}_1}(x_1) m_{\mathsf{y}_1 \to \mathsf{x}_1}(x_1) \xi_{\mathsf{x}_1, \mathsf{x}_2}(x_1, x_2) \\ &= \sum_{x_1} p_{\mathsf{x}_1}(x_1) p_{\mathsf{y}_1|\mathsf{x}_1}(\check{y}_1 | x_1) \frac{p_{\mathsf{x}_1, \mathsf{x}_2}(x_1, x_2)}{p_{\mathsf{x}_1}(x_1) p_{\mathsf{x}_2}(x_2)} = p_{\mathsf{y}_1|\mathsf{x}_2}(\check{y}_1 | x_2) \end{aligned} \tag{2.6}$$

The general forward message $m_{\mathsf{x}_i \to \mathsf{x}_{i+1}}(x_{i+1})$ depends on $m_{\mathsf{y}_i \to \mathsf{x}_i}(x_i)$,$m_{\mathsf{x}_{i-1} \to \mathsf{x}_i}(x_i)$, $\eta_{\mathsf{x}_i}(x_i)$, and $\xi_{\mathsf{x}_i, \mathsf{x}_{i+1}}(x_i, x_{i+1})$.

$$\begin{aligned} m_{\mathsf{x}_i \to \mathsf{x}_{i+1}}(x_{i+1}) &= \sum_{x_i} \eta_{\mathsf{x}_i}(x_i) m_{\mathsf{y}_i \to \mathsf{x}_i}(x_i) m_{\mathsf{x}_{i-1} \to \mathsf{x}_i}(x_i) \xi_{\mathsf{x}_i, \mathsf{x}_{i+1}}(x_i, x_{i+1}) \\ &= \sum_{x_i} p_{\mathsf{x}_i}(x_i) p_{\mathsf{y}_i|\mathsf{x}_i}(\check{y}_i | x_i) p_{\mathsf{y}_1^{i-1}|\mathsf{x}_i}(\check{y}_1, \ldots, \check{y}_{i-1} | x_i) \frac{p_{\mathsf{x}_i, \mathsf{x}_{i+1}}(x_i, x_{i+1})}{p_{\mathsf{x}_i}(x_i) p_{\mathsf{x}_{i+1}}(x_{i+1})} \\ &= p_{\mathsf{y}_1^i|\mathsf{x}_{i+1}}(\check{y}_1^i | x_{i+1}) \end{aligned} \tag{2.7}$$

Using these forward message we will infer the posterior:

$$\begin{aligned} p_{\mathsf{x}_n|\mathsf{y}_1^{n-1}}(x_n | \check{y}_1^{n-1}) &\propto p_{\mathsf{x}_n, \mathsf{y}_1^{n-1}}(x_n, \check{y}_1^{n-1}) \\ &= m_{\mathsf{x}_{n-1} \to \mathsf{x}_n}(x_n) \eta_{\mathsf{x}_n}(x_n) \end{aligned} \tag{2.8}$$

The computational complexity of the forward messages ($m_{\mathsf{x}_i \to \mathsf{x}_{i+1}}(x_{i+1})$), is $\mathcal{O}(|\mathcal{X}|^2)$, where $|\mathcal{X}|$ is the size of the state space. Since we are calculating $n-1$ forward messages to infer the posterior, then the computational complexity of the sum-product algorithm is $\mathcal{O}(n|\mathcal{X}|^2)$. Therefore, the computational complexity of the sum-product algorithm is proportional to the size of the state space and the number of time steps.

In terms of communication complexity, the messages in the sum-product algorithm consist of the sufficient statistics, which are either the emission or transition probabilities. Therefore, the communication complexity is $\mathcal{O}(|\mathcal{X}|)$ for each message. However, an alternative approach can be used at the observation node by sending the observation directly to the hidden node, instead of the emission probability. In this case, the communication complexity for this message is $\mathcal{O}(|\mathcal{Y}|)$, where $|\mathcal{Y}|$ is the size of the observation space. Therefore, the communication complexity is proportional to the size of the state or observation space, depending on the message format used.

As high-dimensional data becomes increasingly prevalent in modern applications, the size of both observation and state spaces can become prohibitively large. This can pose significant challenges for inference algorithms, leading to issues with both computational and communication complexity that can limit their effectiveness and efficiency.

To address these challenges, it is essential to develop a better understanding of quantitative metrics that can help us determine which fragments of information are most important for specific inference tasks. By identifying these important pieces of information, we can make more informed decisions about which parts of the data to prioritize, and which parts can be safely dropped or simplified.

This approach can help to reduce the computational and communication complexity of inference algorithms, while still maintaining a high degree of accuracy and effectiveness. By striking the right balance between information content and computational/communication cost, we can ensure that our models are both efficient and effective, even in the face of high-dimensional data and complex inference tasks.

In the next section, we will introduce the Divergence Transfer Matrix (DTM) [40], which characterizes the modal decomposition of the joint distribution of two discrete random variables.

2.3 Joint Distribution Modal Decomposition

DTM characterizes the modal decomposition of the joint distribution of two discrete random variables. This decomposition can be used to identify the most relevant and informative features of the data, allowing us to focus our attention on the parts of the data that are most important for specific inference tasks. The DTM is defined as:

$$B_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2) \triangleq \frac{p_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2)}{\sqrt{p_{\mathsf{x}_1}(x_1)}\sqrt{p_{\mathsf{x}_2}(x_2)}} \tag{2.9}$$

where $p_{\mathsf{x}_1}(x_1) > 0$ and $p_{\mathsf{x}_2}(x_2) > 0$ for all $x_1, x_2 \in \mathcal{X}$. The $\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2}$ matrix can also be expressed using matrix notation as:

$$\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2} = \left[\sqrt{\mathbf{P}_{\mathsf{x}_2}}\right]^{-1} \mathbf{P}_{\mathsf{x}_2,\mathsf{x}_1} \left[\sqrt{\mathbf{P}_{\mathsf{x}_1}}\right]^{-1} \tag{2.10}$$

where $[\sqrt{\mathbf{P}_{\mathsf{x}_1}}]^{-1}$ denotes a $|\mathcal{X}| \times |\mathcal{X}|$ diagonal matrix whose xth diagonal entry is $\sqrt{p_{x_1}(x_1)}$, where $[\sqrt{\mathbf{P}_{\mathsf{x}_2}}]^{-1}$ denotes a $|\mathcal{X}| \times |\mathcal{X}|$ diagonal matrix whose xth diagonal entry is $\sqrt{p_{x_2}(x_2)}$, and where $\mathbf{P}_{\mathsf{x}_2,\mathsf{x}_1}$ denotes a $|\mathcal{X}| \times |\mathcal{X}|$ matrix whose (x_2, x_1)th entry is $p_{\mathsf{x}_2,\mathsf{x}_1}(x_2, x_1)$.

The Singular Value Decomposition (SVD) of the $\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2}$ takes the form:

$$\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2} = \sum_{i=0}^{|\mathcal{X}|-1} \sigma_i \boldsymbol{\psi}_i^{\mathsf{x}_2} (\boldsymbol{\psi}_i^{\mathsf{x}_1})^{\mathsf{T}} \tag{2.11}$$

where σ_i denotes the ith singular values and where $\boldsymbol{\psi}_i^{\mathsf{x}_2}$ and $\boldsymbol{\psi}_i^{\mathsf{x}_1}$ are the corresponding left and right singular vectors. The singular values are ordered according to

$$\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{|\mathcal{X}|} \tag{2.12}$$

The following proposition establishes that $\mathbf{B}$ is a contractive operator.

Proposition 2.1. *For* $\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2}$ *defined via* (2.10) *we have*

$$\left\|\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2}\right\|_s = 1 \tag{2.13}$$

where $||.||_s$ *denotes the spectral norm. Moreover, the left and right singular vectors* $\boldsymbol{\psi}_0^{\mathsf{x}_2}$ *and* $\boldsymbol{\psi}_0^{\mathsf{x}_1}$ *associated with singular value* $\sigma_0 = 1$ *have elements*

$$\psi_0^{\mathsf{x}_1}(x_1) \triangleq \sqrt{p_{\mathsf{x}_1}(x_1)} \tag{2.14}$$

$$\psi_0^{\mathsf{x}_2}(x_2) \triangleq \sqrt{p_{\mathsf{x}_2}(x_2)} \tag{2.15}$$

Proof. A proof is provided in Appendix II-A in [40] □

Using the second part of Proposition 2.1, it follows immediately that $\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2}$ is an equivalent representation of $\mathbf{P}_{\mathsf{x}_2,\mathsf{x}_1}$. Indeed, given $\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2}$, we can compute the singular vectors $\boldsymbol{\psi}_0^{\mathsf{X}_1}$ and $\boldsymbol{\psi}_0^{\mathsf{X}_2}$, from which we obtain the marginal probabilities $\mathbf{p}_{\mathsf{x}_1}$ and $\mathbf{p}_{\mathsf{x}_2}$. In addition we can find the joint distribution:

$$\left[\sqrt{\mathbf{P}_{\mathsf{x}_2}}\right]\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2}\left[\sqrt{\mathbf{P}_{\mathsf{x}_1}}\right] = \left[\sqrt{\mathbf{P}_{\mathsf{x}_2}}\right]\left[\sqrt{\mathbf{P}_{\mathsf{x}_2}}\right]^{-1}\mathbf{P}_{\mathsf{x}_2,\mathsf{x}_1}\left[\sqrt{\mathbf{P}_{\mathsf{x}_1}}\right]^{-1}\left[\sqrt{\mathbf{P}_{\mathsf{x}_1}}\right] = \mathbf{P}_{\mathsf{x}_2,\mathsf{x}_1}$$

The SVD of (2.11) provides a key expansion of the joint distribution $p_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2)$ as shown in the following result.

Proposition 2.2. *Let* $\mathcal{X}$ *denotes a finite alphabet. Then for any joint distribution* $\mathbf{P}_{\mathsf{x}_1,\mathsf{x}_2}$*, there exist features* $f_i^* : \mathcal{X} \to \mathbb{R}$ *and* $g_i^* : \mathcal{X} \to \mathbb{R}$*, for* $i = 1, \cdots, |\mathcal{X}| - 1$*, such that*

$$p_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2) = p_{x_1}(x_1)p_{x_2}(x_2)\left[1 + \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i f_i^*(x_1)g_i^*(x_2)\right] \tag{2.16}$$

where $\sigma_1, \cdots, \sigma_{|\mathcal{X}|-1}$ *are defined as in* (2.11) *and where*

$$\mathbb{E}[f_i^*(X_1)] = 0, \quad i \in \{1, \cdots, |\mathcal{X}| - 1\} \tag{2.17a}$$

$$\mathbb{E}[g_i^*(X_2)] = 0, \quad i \in \{1, \cdots, |\mathcal{X}| - 1\} \tag{2.17b}$$

$$\mathbb{E}[f_i^*(X_1)f_j^*(X_1)] = \mathbb{1}_{i=j}, \quad i, j \in \ 1, \cdots, |\mathcal{X}| - 1\} \tag{2.17c}$$

$$\mathbb{E}[g_i^*(X_2)g_j^*(X_2)] = \mathbb{1}_{i=j}, \quad i, j \in \ 1, \cdots, |\mathcal{X}| - 1\} \tag{2.17d}$$

Moreover, f_i^* *and* g_i^* *are related to the singular vector in* (2.11) *according to*

$$f_i^*(x_1) \triangleq \frac{\psi_i^{\mathsf{x}_1}(x_1)}{\sqrt{p_{\mathsf{x}_1}(x_1)}} \tag{2.18}$$

$$g_i^*(x_2) \triangleq \frac{\psi_i^{\mathsf{x}_2}(x_2)}{\sqrt{p_{\mathsf{x}_2}(x_2)}} \tag{2.19}$$

Proof.

$$\begin{aligned}
B_{\mathsf{x}_1,\mathsf{x}_2}(x_1,x_2) &= \frac{p_{\mathsf{x}_1,\mathsf{x}_2}(x_1,x_2)}{\sqrt{p_{\mathsf{x}_1}(x_1)}\sqrt{p_{\mathsf{x}_2}(x_2)}} \\
&= \sqrt{p_{x_1}(x_1)}\sqrt{p_{x_2}(x_2)} + \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i \psi_i^{\mathsf{x}_1}(x_1)\psi_i^{\mathsf{x}_2}(x_2) \\
&= \sqrt{p_{x_1}(x_1)}\sqrt{p_{x_2}(x_2)} + \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i \sqrt{p_{x_1}(x_1)} f_i^*(x_1)\sqrt{p_{x_1}(x_1)} g_i^*(x_2) \\
&= \sqrt{p_{x_1}(x_1)}\sqrt{p_{x_2}(x_2)} \left[1 + \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i f_i^*(x_1) g_i^*(x_2)\right] \\
\implies p_{\mathsf{x}_1,\mathsf{x}_2}(x_1,x_2) &= p_{x_1}(x_1)p_{x_2}(x_2)\left[1 + \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i f_i^*(x_1) g_i^*(x_2)\right]
\end{aligned}$$

□

Proposition 2.2 demonstrates a method of decomposing the joint distribution of two discrete random variables through the use of the singular values and singular vectors of the $\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2}$. This decomposition provides a level of flexibility in determining the amount of information we wish to capture about the joint distribution. The ability to control the level of information can be particularly useful in scenarios where computational and communication complexity is a concern. For instance, in cases where the size of $\mathcal{X}$ is small, all of the singular values and their associated singular vectors can be utilized. However, when $\mathcal{X}$ is large, fewer singular values and their associated singular vectors will need to be employed to stay within computational and communication limitations.

In addition to its usefulness in dealing with joint distributions, Proposition 2.2

can also be extended to address conditional distributions. By applying the same decomposition technique, the results can be extended to this case as well.

$$p_{\mathsf{x}_1|\mathsf{x}_2}(x_1|x_2) = p_{x_1}(x_1)\left[1 + \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i f_i^*(x_1) g_i^*(x_2)\right] \tag{2.20}$$

$$p_{\mathsf{x}_2|\mathsf{x}_1}(x_2|x_1) = p_{x_2}(x_2)\left[1 + \sum_{i=1}^{|\mathcal{X}|-1} \sigma_i f_i^*(x_1) g_i^*(x_2)\right] \tag{2.21}$$

The Canonical Dependence Matrix (CDM) is defined at the DTM matrix without the zeroth mode. Therefore, the CDM is defined as

$$\hat{B}_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2) \triangleq \frac{p_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2) - p_{\mathsf{x}_1}(x_1) p_{\mathsf{x}_2}(x_2)}{\sqrt{p_{\mathsf{x}_1}(x_1)}\sqrt{p_{\mathsf{x}_2}(x_2)}} \tag{2.22}$$

The significance of the CDM lies in its characteristic of having the largest singular value ($\sigma_1 \leq 1$), a feature that differs from the DTM, which consistently has the leading singular value at 1. The benefits of this distinction become apparent as we delve into the discussion of the computational complexity constrained information exchange message passing algorithm in Chapter 3.

In the upcoming section, we aim to introduce the KL divergence under the weakly dependent regime. This measure can be utilized to assess the effect of either limiting the exchange of information between nodes or in other words discarding a portion of the available information. By leveraging the KL divergence, we can quantify the degree of deviation that arises due to the constraints imposed on the system.

2.4 Kullback-Leibler (KL) Divergence under Weakly Dependent Regime

Kullback-Leibler (KL) divergence, also known as relative entropy, is a fundamental concept in information theory and statistics. It quantifies the difference between two probability distributions. The KL divergence is defined as [41]

$$D(\mathbf{p}||\mathbf{q}) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right) \tag{2.23}$$

where $\mathbf{p}, \mathbf{q}$ are discrete probability distribution.

KL divergence is non-negative and asymmetric, meaning that $D(\mathbf{p}||\mathbf{q}) \neq D(\mathbf{q}||\mathbf{p})$. KL divergence finds various applications in fields such as machine learning, information theory, and natural language processing. However, there no systematic approach for finding the optimum solution in general. The main source of difficulty is that the KL divergence is not a metric in the space of probability distributions [42]. In fact, the collection of distributions in general, form a manifold, which invalidates interpreting the KL divergence as a distance between distributions. A natural way to simplify this general scenario is to restrict our attention to a local neighborhood of distributions, in which the manifold behaves like a Euclidean space and the KL divergence behaves like the Euclidean metric [42]. This lead us to define the weakly dependent random variables.

Definition 2.3. *Let* x *and* y *be defined over alphabet* $\mathcal{X}$ *and* $\mathcal{Y}$*, respectively, and distributed according to* $\mathbf{P}_{\mathsf{x},\mathsf{y}} \in \mathcal{P}^{\mathcal{X}\times\mathcal{Y}}$ *is the usual restriction of the simplex to distribution with strictly positive marginals. Then* x *and* y *are* ϵ*-dependent if there exists an* $\epsilon > 0$ *such that*

$$\mathbf{P}_{\mathsf{x},\mathsf{y}} \in \mathcal{N}_{\epsilon}^{\mathcal{X}\times\mathcal{Y}}(\mathbf{p}_{\mathsf{x}}\mathbf{p}_{\mathsf{y}}) = \{D_{\mathcal{X}^2}(\mathbf{P}_{\mathsf{x},\mathsf{y}}||\mathbf{p}_{\mathsf{x}}\mathbf{p}_{\mathsf{y}}) = \sum_{x,y}\frac{(p_{\mathsf{x},\mathsf{y}}(x,y) - p_{\mathsf{x}}(x)p_{\mathsf{y}}(y))^2}{p_{\mathsf{x}}(x)p_{\mathsf{y}}(y)} \leq \epsilon^2\} \tag{2.24}$$

where $\mathbf{p}_{\mathsf{x}}$ *and* $\mathbf{p}_{\mathsf{y}}$ *are the marginal distributions. This definition also implies*

$$\mathbf{p}_{\mathsf{x}|\mathsf{y}}(.|\check{y}) \in \mathcal{N}_{\epsilon}^{\mathcal{X}}(\mathbf{p}_{\mathsf{x}}) = \{D_{\mathcal{X}^2}(\mathbf{p}_{\mathsf{x}|\mathsf{y}}(.|\check{y})||\mathbf{p}_{\mathsf{x}}) = \sum_{x}\frac{(p_{\mathsf{x}|\mathsf{y}}(x|\check{y}) - p_{\mathsf{x}}(x))^2}{p_{\mathsf{x}}(x)} \leq \epsilon^2\} \tag{2.25}$$

Another way to define the ϵ-dependence is by defining the information vector

$$\phi^{(\mathsf{x}|\mathsf{y})}(x|\check{y}) = \frac{p_{\mathsf{x}|\mathsf{y}}(x|\check{y}) - p_{\mathsf{x}}(x)}{\epsilon\sqrt{p_{\mathsf{x}}(x)}}; \quad ||\boldsymbol{\phi}^{(\mathsf{x}|\mathsf{y})}||_2^2 \leq 1 \tag{2.26}$$

Lemma 2.4. *For a given* $\mathbf{p}_{\mathsf{x}}$ *and* $\epsilon > 0$*, let* $\mathbf{p}_1, \mathbf{p}_2 \in \mathcal{N}_{\epsilon}^{\mathcal{X}}(\mathbf{p}_{\mathsf{x}})$ *be arbitrary, and let* $\boldsymbol{\phi}_1$ *and* $\boldsymbol{\phi}_2$ *denote the corresponding information vectors, respectively. Then*

$$D(\mathbf{p}_1||\mathbf{p}_2) \triangleq \sum_{x\in\mathcal{X}} p_1(x) \log \frac{p_1(x)}{p_2(x)} = \frac{\epsilon^2}{2}||\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2||_2^2 + o(\epsilon^2) \tag{2.27}$$

Proof: A proof is provided in A.1

Leveraging the mathematical result presented in lemma 2.4, we can reduce the computational complexity involved in calculating the KL divergence. Specifically, we can simplify the computation to the evaluation of the Euclidean norm between two information vectors. This simplification can be highly beneficial when optimizing the KL divergence. This approximation is commonly employed in practical scenarios where the output shows weak dependency on the input [43, 44]. This occurs in situations of weak supervised learning where the training data is assigned more generalized labels instead of precise ones. The adoption of weak supervised learning is gaining popularity due to the substantial costs tied to labeling large training datasets [45]. Fig. 2-2 illustrates the ϵ value as a function of the number of tags in penn treebank dataset [46]. As it can seen in Fig. 2-2, as the number of tags decreases due to grouping the different tags together the epsilon value gets smaller and this is because the data and labels become weakly dependent.

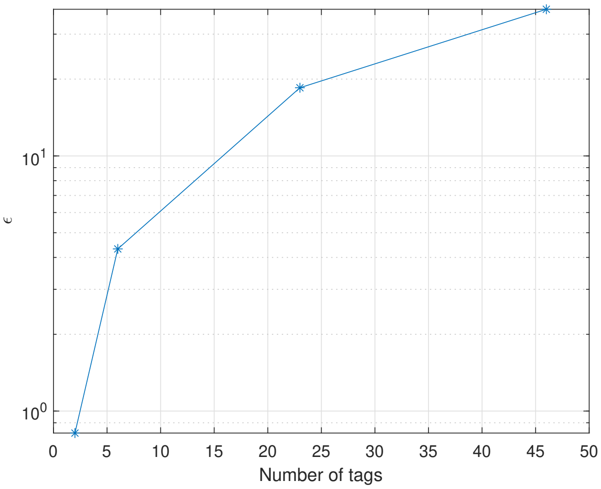


Figure 2-2: ϵ value as a function of the number of tags of the penn treebank dataset

In the next section, we plan to explore the concept of weakly correlated variables and the DTM/CDM matrix in the context of Hidden Markov Models (HMMs). By incorporating these ideas, we aim to establish a relationship between the observation and hidden nodes of an HMM.

2.5 Local Geometry of Attribute Variables in Hidden Markov Model

The development and analysis of the constrained information exchange message passing algorithm involve the integration of two important concepts: weakly correlated variables and the DTM/CDM. These elements are instrumental in understanding the underlying system's dynamics and are crucial in designing and analyzing the algorithm.

When designing the algorithm, it is essential to consider the local geometries present among the attributes of the Hidden Markov Model (HMM). These local geometries reflect the relationships and dependencies between different attributes such as observation variables and hidden states. By taking into account these local geometries, the algorithm can effectively propagate and exchange information, constrained by the structure of the HMM. This ensures that the algorithm accurately captures the intricate dynamics and dependencies within the system.

Specifically, in the context of the algorithm, the variables x_i and y_i are considered weakly correlated and this implies we will have the following information vector

$$\phi^{(\mathsf{x}_i|\mathsf{y}_i)}(x_i|\check{y}_i) = \frac{p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i) - p_{\mathsf{x}_i}(x_i)}{\epsilon\sqrt{p_{\mathsf{x}_i}(x_i)}} \tag{2.28}$$

This will only characterize the relationship between x_i and y_i but lays the foundation for the relationship between x_{i+1} and y_i as shown in the following lemma.

Lemma 2.5. *For a given* $\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}}$,$\mathbf{p}_{\mathsf{x}_i}$,$\mathbf{p}_{\mathsf{x}_{i+1}}$ *and* $\epsilon > 0$, *let* $\mathbf{p}_{x_i|y_i}(.|\check{y}_i) \in \mathcal{N}_\epsilon^{\mathcal{X}}(\mathbf{p}_{\mathsf{x}_i})$ *and*

$\mathbf{p}_{x_{i+1}|y_i}(.|\check{y}_i) \in \mathcal{N}_\epsilon^{\mathcal{X}}(\mathbf{p}_{\mathsf{x}_{i+1}})$ be arbitrary, and let $\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}(.|\check{y}_i)$ and $\boldsymbol{\phi}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)}(.|\check{y}_i)$ denote the corresponding information vectors, respectively. Then

$$\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}(.|\check{y}_i) = \boldsymbol{\phi}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)}(.|\check{y}_i) \tag{2.29}$$

Proof.

$$\begin{aligned}
&\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}(.|\check{y}_i) \\
&= \sum_{\mathsf{x}_i \in \mathcal{X}} \frac{p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i, x_{i+1})}{\sqrt{p_{\mathsf{x}_i}(x_i)}\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}} \frac{p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i) - p_{\mathsf{x}_i}(x_i)}{\epsilon\sqrt{p_{\mathsf{x}_i}(x_i)}} \\
&= \frac{1}{\epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}} \sum_{\mathsf{x}_i \in \mathcal{X}} \frac{p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i, x_{i+1}) p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i)}{p_{\mathsf{x}_i}(x_i)} - \frac{p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i, x_{i+1}) p_{\mathsf{x}_i}(x_i)}{\mathsf{x}_i(x_i)} \\
&= \frac{1}{\epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}} \sum_{\mathsf{x}_i \in \mathcal{X}} p_{\mathsf{x}_{i+1}|\mathsf{x}_i}(x_{i+1}|x_i) p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i) - \frac{1}{\epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}} \sum_{\mathsf{x}_i \in \mathcal{X}} p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i, x_{i+1}) \\
&= \frac{1}{\epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}} \Big(p_{\mathsf{x}_{i+1}|\mathsf{y}_i}(x_{i+1}|\check{y}_i) - p_{\mathsf{x}_{i+1}}(x_{i+1})\Big) \\
&= \boldsymbol{\phi}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)}(.|\check{y}_i)
\end{aligned}$$

□

By looking at lemma 2.5 we can see that we can relate the local geometry of $p_{x_i|y_i}$ and $p_{x_{i+1}|y_i}$ using the DTM matrix. This result can be extended to the CDM as shown in the following lemma.

Lemma 2.6. *For a given $\hat{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}}$,$\mathbf{p}_{\mathsf{x}_i}$,$\mathbf{p}_{\mathsf{x}_{i+1}}$ and $\epsilon > 0$, let $\mathbf{p}_{x_i|y_i}(.|\check{y}_i) \in \mathcal{N}_\epsilon^{\mathcal{X}}(\mathbf{p}_{\mathsf{x}_i})$ and $\mathbf{p}_{x_{i+1}|y_i}(.|\check{y}_i) \in \mathcal{N}_\epsilon^{\mathcal{X}}(\mathbf{p}_{\mathsf{x}_{i+1}})$ be arbitrary, and let $\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}(.|\check{y}_i)$ and $\boldsymbol{\phi}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)}(.|\check{y}_i)$ denote the corresponding information vectors, respectively. Then*

$$\hat{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}(.|\check{y}_i) = \boldsymbol{\phi}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)}(.|\check{y}_i) \tag{2.30}$$

Proof.

$$\begin{aligned}
&\hat{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}(.|\check{y}_i)\\
&=\sum_{\mathsf{x}_i\in\mathcal{X}}\frac{p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i,x_{i+1})-p_{\mathsf{x}_i}(x_i)p_{\mathsf{x}_{i+1}}(x_{i+1})}{\sqrt{p_{\mathsf{x}_i}(x_i)}\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}}\frac{p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i)-p_{\mathsf{x}_i}(x_i)}{\epsilon\sqrt{p_{\mathsf{x}_i}(x_i)}}\\
&=\frac{1}{\epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}}\sum_{\mathsf{x}_i\in\mathcal{X}}\frac{p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i,x_{i+1})p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i)}{p_{\mathsf{x}_i}(x_i)}-\frac{p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i,x_{i+1})p_{\mathsf{x}_i}(x_i)}{p_{\mathsf{x}_i}(x_i)}\\
&-\frac{\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}}{\epsilon}\sum_{\mathsf{x}_i\in\mathcal{X}}p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i)-p_{\mathsf{x}_i}(x_i)\\
&=\frac{1}{\epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}}\sum_{\mathsf{x}_i\in\mathcal{X}}p_{\mathsf{x}_{i+1}|\mathsf{x}_i}(x_{i+1}|x_i)p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i)-\frac{1}{\epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}}\sum_{\mathsf{x}_i\in\mathcal{X}}p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i,x_{i+1})\\
&=\frac{1}{\epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}}\Big(p_{\mathsf{x}_{i+1}|\mathsf{y}_i}(x_{i+1}|\check{y}_i)-p_{\mathsf{x}_{i+1}}(x_{i+1})\Big)\\
&=\boldsymbol{\phi}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)}(.|\check{y}_i)
\end{aligned}$$

□

This pave the way to for a more general lemma which relates different attribute variables of the HMM.

Lemma 2.7. *For a given* $\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2},\cdots,\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}}$, $\mathbf{p}_{\mathsf{x}_1},\cdots,\mathbf{p}_{\mathsf{x}_i}$, *and* $\epsilon>0$, *let* $\mathbf{p}_{x_j|y_j}(.|\check{y}_j)\in\mathcal{N}_\epsilon^{\mathcal{X}}(\mathbf{p}_{\mathsf{x}_j})$ *and let* $\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}(.|\check{y}_j)$ *be the corresponding information vector for* $j\in\{1,\cdots,i\}$ *then*

$$\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_1^i)}(.|\check{y}_1^i)=\sum_{j=1}^{i}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_j)}(.|\check{y}_j)+o(\epsilon)\tag{2.31}$$

Proof. A proof is provided in A.2 □

By looking at Lemma 2.7 we can see that the information vector $\phi^{(\mathsf{x}_i|\mathsf{y}_1^i)}$ is a superposition of other information vectors with different observation. This result can be extended to the case of the CDM matrix. This property will be used to modify the sum-product algorithm in terms of the *information vector* and the *DTM/CDM matrix.*

2.6 Sum-Product Algorithm under Weakly Dependent Regime

The modified sum-product algorithm in terms of the *information vector* and the *DTM/CDM matrix* is of great importance because it will give us the flexibility of varying the computational and communication complexity of the inference algorithm and striking the right balance between information content and computational cost, we can ensure that our models are both efficient and effective, even in the face of high-dimensional data and complex inference tasks.

2.6.1 Modified Sum Product Algorithm Using DTM

The modified sum-product algorithm forward message $m_{\mathsf{y}_i \to \mathsf{x}_i}(x_i)$ is the information vector associated with these two nodes:

$$m_{\mathsf{y}_i \to \mathsf{x}_i}(x_i) = \phi^{(\mathsf{x}_i|\mathsf{y}_i)}(x_i|\check{y}_i)$$

The forward message $m_{\mathsf{x}_1 \to \mathsf{x}_2}(x_2)$ is the message from $\mathbf{m}_{\mathsf{y}_1 \to \mathsf{x}_1}$ multiplied by the DTM matrix $\mathbf{B}_{\mathsf{x}_1,\mathsf{x}_2}$ utilizing lemma 2.5

$$\begin{aligned}
m_{\mathsf{x}_1 \to \mathsf{x}_2}(x_2) &= \sum_{x_1} \frac{p_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2)}{\sqrt{p_{\mathsf{x}_1}(x_1)}\sqrt{p_{\mathsf{x}_2}(x_2)}} m_{\mathsf{y}_1 \to \mathsf{x}_1}(x_1) \\
&= \sum_{x_1} \frac{p_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2)}{\sqrt{p_{\mathsf{x}_1}(x_1)}\sqrt{p_{\mathsf{x}_2}(x_2)}} \phi^{(\mathsf{x}_1|\mathsf{y}_1)}(x_1|\check{y}_1) \\
&= \phi^{(\mathsf{x}_2|\mathsf{y}_1)}(x_2|\check{y}_1)
\end{aligned}$$

The forward message $m_{\mathsf{x}_2 \to \mathsf{x}_3}(x_3)$ is the message from $m_{\mathsf{x}_1 \to \mathsf{x}_2}(x_2)$ and $m_{\mathsf{y}_1 \to \mathsf{x}_1}(x_1)$ and multiplied by the DTM matrix $\mathbf{B}_{\mathsf{x}_2,\mathsf{x}_3}$ utilizing lemma 2.7

$$\begin{aligned}
m_{\mathsf{x}_2 \to \mathsf{x}_3}(x_3) &= \sum_{x_2} \frac{p_{\mathsf{x}_2,\mathsf{x}_3}(x_2, x_3)}{\sqrt{p_{\mathsf{x}_2}(x_2)}\sqrt{p_{\mathsf{x}_3}(x_3)}} (m_{\mathsf{y}_2 \to \mathsf{x}_2}(x_2) + m_{\mathsf{x}_1 \to \mathsf{x}_2}(x_2)) \\
&= \sum_{x_2} \frac{p_{\mathsf{x}_2,\mathsf{x}_3}(x_2, x_3)}{\sqrt{p_{\mathsf{x}_2}(x_2)}\sqrt{p_{\mathsf{x}_3}(x_3)}} (\phi^{(\mathsf{x}_2|\mathsf{y}_1)}(x_2|\check{y}_1) + \phi^{(\mathsf{x}_2|\mathsf{y}_2)}(x_2|\check{y}_2))
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{x_2} \frac{p_{\mathsf{x}_2,\mathsf{x}_3}(x_2, x_3)}{\sqrt{p_{\mathsf{x}_2}(x_2)}\sqrt{p_{\mathsf{x}_3}(x_3)}} (\phi^{(\mathsf{x}_2|\mathsf{y}_1^2)}(x_2|\check{y}_1^2)) \\
&= \phi^{(\mathsf{x}_3|\mathsf{y}_1^2)}(x_3|\check{y}_1^2)
\end{aligned}
$$

Therefore, the general forward message $m_{\mathsf{x}_i \to \mathsf{x}_{i+1}}(x_{i+1})$ is

$$
\begin{aligned}
m_{\mathsf{x}_i \to \mathsf{x}_{i+1}}(x_{i+1}) &= \sum_{x_i} \frac{p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i, x_{i+1})}{\sqrt{p_{\mathsf{x}_i}(x_i)}\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}} (m_{\mathsf{y}_i \to \mathsf{x}_i}(x_i) + m_{\mathsf{x}_{i-1} \to \mathsf{x}_i}(x_i)) \\
&= \sum_{x_i} \frac{p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i, x_{i+1})}{\sqrt{p_{\mathsf{x}_i}(x_i)}\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}} (\phi^{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i) + \phi^{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(x_i|\check{y}_1^{i-1})) \\
&= \phi^{(\mathsf{x}_{i+1}|\mathsf{y}_1^i)}(x_{i+1}|\check{y}_1^i)
\end{aligned}
$$

Using these forward messages we will infer the posterior:

$$
p_{\mathsf{x}_{i+1}|y_1^i}(x_{i+1}|\check{y}_1^i) = p_{\mathsf{x}_{i+1}}(x_{i+1}) + \epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})} m_{\mathsf{x}_i \to \mathsf{x}_{i+1}}(x_{i+1})
$$

2.6.2 Modified Sum Product Algorithm Using CDM

The modified sum-product algorithm forward message $m_{\mathsf{y}_i \to \mathsf{x}_i}(x_i)$ is the information vector associated with these two nodes:

$$
m_{\mathsf{y}_i \to \mathsf{x}_i}(x_i) = \phi^{(\mathsf{x}_i|\mathsf{y}_i)}(x_i|\check{y}_i)
$$

The forward message $m_{\mathsf{x}_1 \to \mathsf{x}_2}(x_2)$ is the message from $\mathbf{m}_{\mathsf{y}_1 \to \mathsf{x}_1}$ multiplied by the CDM matrix $\hat{\mathbf{B}}_{\mathsf{x}_1,\mathsf{x}_2}$ utilizing lemma 2.6

$$
\begin{aligned}
m_{\mathsf{x}_1 \to \mathsf{x}_2}(x_2) &= \sum_{x_1} \frac{p_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2) - p_{\mathsf{x}_1}(x_1)p_{\mathsf{x}_2}(x_2)}{\sqrt{p_{\mathsf{x}_1}(x_1)}\sqrt{p_{\mathsf{x}_2}(x_2)}} m_{\mathsf{y}_1 \to \mathsf{x}_1}(x_1) \\
&= \sum_{x_1} \frac{p_{\mathsf{x}_1,\mathsf{x}_2}(x_1, x_2) - p_{\mathsf{x}_1}(x_1)p_{\mathsf{x}_2}(x_2)}{\sqrt{p_{\mathsf{x}_1}(x_1)}\sqrt{p_{\mathsf{x}_2}(x_2)}} \phi^{(\mathsf{x}_1|\mathsf{y}_1)}(x_1|\check{y}_1) \\
&= \phi^{(\mathsf{x}_2|\mathsf{y}_1)}(x_2|\check{y}_1)
\end{aligned}
$$

The forward message $m_{\mathsf{x}_2\to\mathsf{x}_3}(x_3)$ is the message from $m_{\mathsf{x}_1\to\mathsf{x}_2}(x_2)$ and $m_{\mathsf{y}_1\to\mathsf{x}_1}(x_1)$ and multiplied by the CDM matrix $\hat{\mathbf{B}}_{\mathsf{x}_2,\mathsf{x}_3}$ utilizing lemma 2.7

$$\begin{aligned}
m_{\mathsf{x}_2\to\mathsf{x}_3}(x_3) &= \sum_{x_2} \frac{p_{\mathsf{x}_2,\mathsf{x}_3}(x_2,x_3) - p_{\mathsf{x}_2}(x_2)p_{\mathsf{x}_3}(x_3)}{\sqrt{p_{\mathsf{x}_2}(x_2)}\sqrt{p_{\mathsf{x}_3}(x_3)}}(m_{\mathsf{y}_2\to\mathsf{x}_2}(x_2) + m_{\mathsf{x}_1\to\mathsf{x}_2}(x_2)) \\
&= \sum_{x_2} \frac{p_{\mathsf{x}_2,\mathsf{x}_3}(x_2,x_3) - p_{\mathsf{x}_2}(x_2)p_{\mathsf{x}_3}(x_3)}{\sqrt{p_{\mathsf{x}_2}(x_2)}\sqrt{p_{\mathsf{x}_3}(x_3)}}(\phi^{(\mathsf{x}_2|\mathsf{y}_1)}(x_2|\check{y}_1) + \phi^{(\mathsf{x}_2|\mathsf{y}_2)}(x_2|\check{y}_2)) \\
&= \sum_{x_2} \frac{p_{\mathsf{x}_2,\mathsf{x}_3}(x_2,x_3) - p_{\mathsf{x}_2}(x_2)p_{\mathsf{x}_3}(x_3)}{\sqrt{p_{\mathsf{x}_2}(x_2)}\sqrt{p_{\mathsf{x}_3}(x_3)}}(\phi^{(\mathsf{x}_2|\mathsf{y}_1^2)}(x_2|\check{y}_1^2)) \\
&= \phi^{(\mathsf{x}_3|\mathsf{y}_1^2)}(x_3|\check{y}_1^2)
\end{aligned}$$

Therefore, the general forward message $m_{\mathsf{x}_i\to\mathsf{x}_{i+1}}(x_{i+1})$ is

$$\begin{aligned}
m_{\mathsf{x}_i\to\mathsf{x}_{i+1}}(x_{i+1}) &= \sum_{x_i} \frac{p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i,x_{i+1}) - p_{\mathsf{x}_i}(x_i)p_{\mathsf{x}_{i+1}}(x_{i+1})}{\sqrt{p_{\mathsf{x}_i}(x_i)}\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}}(m_{\mathsf{y}_i\to\mathsf{x}_i}(x_i) + m_{\mathsf{x}_{i-1}\to\mathsf{x}_i}(x_i)) \\
&= \sum_{x_i} \frac{p_{\mathsf{x}_i,\mathsf{x}_{i+1}}(x_i,x_{i+1}) - p_{\mathsf{x}_i}(x_i)p_{\mathsf{x}_{i+1}}(x_{i+1})}{\sqrt{p_{\mathsf{x}_i}(x_i)}\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}}(\phi^{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i) + \phi^{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(x_i|\check{y}_1^{i-1})) \\
&= \phi^{(\mathsf{x}_{i+1}|\mathsf{y}_1^i)}(x_{i+1}|\check{y}_1^i)
\end{aligned}$$

Using these forward messages we will infer the posterior:

$$p_{\mathsf{x}_{i+1}|y_1^i}(x_{i+1}|\check{y}_1^i) = p_{\mathsf{x}_{i+1}}(x_{i+1}) + \epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}m_{\mathsf{x}_i\to\mathsf{x}_{i+1}}(x_{i+1})$$

These forward messages in both cases are not constrained (sufficient statistics) and in the next chapter we will discuss the case of a constrained information exchange message passing algorithm in different posterior inference scenarios.

Chapter 3

Index-Specific Posterior Inference

The sum product algorithm depends heavily on the information exchange between the nodes of the graph and having unconstrained sized messages, resulting in the exact inference of the posterior distribution. Despite its usefulness and ubiquitous applications, high-dimensional data is becoming increasingly prevalent in modern applications, where the size of both observation and state spaces can become prohibitively large. This poses significant challenges for inference algorithms, leading to issues with both computational and communication complexity that can limit their effectiveness and efficiency.

To address these challenges, in the previous chapter, a modified sum product algorithm was introduced in terms of the information vector and DTM/CDM matrix. This modification allows for the determination of which fragments of information are most important for specific inference tasks. By identifying these important pieces of information, informed decisions can be made about which parts of the data to prioritize and which parts can be safely dropped or simplified.

In this chapter, the constrained information exchange message passing algorithm is introduced. The constrained information exchange message passing algorithm constrains the messages one can pass along in the algorithm and deliberately sets them to be less than what is needed to carry sufficient statistics. This formulation forced us to look for "insufficient statistics," which are expected to be the key new concept in the era of big data as it will result in lower computational and communication complexity. This formulation allows for the study of the tradeoff between information content and computational/communication cost, enabling the development of models that are both efficient and effective, even in the face of high-dimensional data and complex inference tasks.

The restrictive nature of the constraints, which limit the exchange of sufficient statistics, inherently complicates the process of posterior inference. This complexity arises from the unique dynamics of the process, where a distinct insufficient statistic is forwarded as a message, depending on the specific scenario of the posterior inference. Consequently, this intricate issue not only amplifies the complexity of the problem but also necessitates the crafting of diverse messages tailored to each unique posterior inference scenario. In order to address these challenges, it becomes crucial to develop and rigorously analyze the algorithm under a broad range of scenarios. By doing so, one can ensure the algorithm's robustness and adaptability, allowing it to effectively handle a multitude of posterior inference conditions, regardless of the constraints imposed on the information exchange.

Posterior inference scenarios are broadly categorized into two types: index-specific and index-free posteriors. This chapter will primarily focus on the introduction of index-specific posterior inference. This sub-category of posterior inference provides a critical examination of specific scenarios where the interest lies either in a singular posterior inference at a specific index (referred to as single index-specific posterior inference) or in multiple posterior inferences at specific indexes (referred to as multiple index-specific posterior inference).

- Single Index-Specific Posterior Inference: In this particular scenario, characterized as one of the simplest scenarios due to the absence of a tradeoff between different objectives, the primary emphasis lies in the posterior inference at a specific index. The objective is to solve a single objective optimization problem, aimed at finding the optimal matrices at each specific node for effectively compressing the messages.

- Multiple Index-Specific Posterior Inference: In this scenario, the primary emphasis lies in performing posterior inference at multiple indexes, which unveils the inherent tradeoff nature of the problem. The optimization plays a vital role in achieving a delicate balance between the different posterior inference objectives. Through careful optimization, the optimal compression matrices at each specific node is obtained and effectively used to compress the messages. This approach enables efficient handling of the tradeoff between the competing objectives, leading to the optimal inference across multiple indexes.

In all these cases, the messages $\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i}$ will be constrained. This constraint is motivated by the fact that observation nodes typically possess low power and limited communication bandwidth compared to the hidden nodes. One can envision the observation nodes as sensor nodes in a wireless sensor network, while the hidden node can be regarded as the central unit (server) responsible for receiving and processing all the data. This discrepancy in power and communication capabilities necessitates the imposition of constraints on the messages $\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i}$.

The KL divergence loss function is optimized to find the matrix compression that results in a minimal error. However, the task of finding the optimal matrix compression for different observations or information vectors can be challenging, particularly when the prediction task is unknown or when there are multiple tasks involved. To address this challenge, the information vectors is defined as rotation-invariant ensemble (RIE) assigning a uniform prior to unknown attributes. This enables the formulation of the problem as a universal matrix selection problem, seeking to identify the optimal compression matrix by minimizing the average KL divergence over the RIE in a weakly dependent regime. This objective is equivalent to minimizing the average mean squared error (MSE) over the RIE. The use of this universal optimal matrix ensures optimal performance on average, without prior knowledge of the prediction task or in the case of multiple tasks. Before delving into discussions regarding different posterior inference scenarios, it is important to define the following lemma:

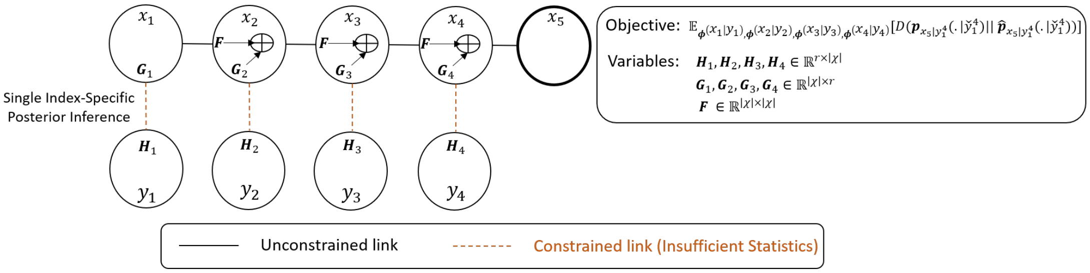
Lemma 3.1. *Let $\mathbf{Z}$ be a $k_1 \times k_2$ spherically symmetric random matrix. Then if $\mathbf{A}_1$ and $\mathbf{A}_2$ are any fixed matrices of compatible dimensions, then*

$$\mathbb{E}\left[||\mathbf{A}_1^T\mathbf{Z}\mathbf{A}_2||_F^2\right] = \frac{1}{k_1k_2}||\mathbf{A}_1||_F^2||\mathbf{A}_2||_F^2\mathbb{E}\left[||\mathbf{Z}||_F^2\right]$$

Proof. The proof is provide in Appendix V-C in [40] □

3.1 Single Index-Specific Posterior Inference

This scenario involves estimating the posterior distribution at a specific index, where a single objective optimization problem is solved to determine the most effective matrices for compressing messages $\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i}$ (as show in Fig. 3-1). Two cases will be studied in this scenario, varying and fixed DTM matrix.



x_1
x_2
x_3
x_4
x_5
G_1
G_2
G_3
G_4
F
F
F
Single Index-Specific Posterior Inference
H_1
H_2
H_3
H_4
y_1
y_2
y_3
y_4
Objective: $\mathbb{E}_{\phi(x_1|y_1),\phi(x_2|y_2),\phi(x_3|y_3),\phi(x_4|y_4)}[D(\boldsymbol{p}_{x_5|y_1^4}(.|\check{y}_1^4)||\hat{\boldsymbol{p}}_{x_5|y_1^4}(.|\check{y}_1^4))]$
Variables: $\boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3, \boldsymbol{H}_4 \in \mathbb{R}^{r\times|\chi|}$
$\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3, \boldsymbol{G}_4 \in \mathbb{R}^{|\chi|\times r}$
$\boldsymbol{F} \in \mathbb{R}^{|\chi|\times|\chi|}$
Unconstrained link
Constrained link (Insufficient Statistics)

Figure 3-1: Constrained Information Exchange Message-Passing Algorithm in Single Index-Specific Posterior Inference

3.1.1 Varying Divergence Transfer Matrix

In this section $\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ is known exactly and the messages are constrained $\hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i} \in \mathbb{R}^r$, where $r < |\mathcal{X}|$.

- at node y_i, pick a matrix $\mathbf{H}_i \in \mathbb{R}^{r\times|\mathcal{X}|}$, to reduce the size of the message ($\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i}$) to r-dimensional message ($\hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i}$) since there is a constrained link between y_i and x_i.

$$\hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i} = \mathbf{H}_i\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i} = \mathbf{H}_i\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$$

- at node x_i, pick a matrix $\mathbf{G}_i \in \mathbb{R}^{|\mathcal{X}|\times r}$ and $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ to calculate the forward message

$$\begin{aligned}\hat{\mathbf{m}}_{\mathsf{x}_i \to \mathsf{x}_{i+1}} &= \mathbf{G}_i \hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i} + \mathbf{F}\hat{\mathbf{m}}_{\mathsf{x}_{i-1} \to \mathsf{x}_i} \\ &= \mathbf{G}_i \mathbf{H}_i \boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)} + \mathbf{F}\hat{\boldsymbol{\phi}}^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})} \\ &= \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)} + \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^{i-1})} \\ &= \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^i)}\end{aligned}$$

- Optimization:

$$\begin{aligned}&\min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}} \left[D(\mathbf{p}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(.|\check{\mathsf{y}}_1^i)||\hat{\mathbf{p}}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(.|\check{\mathsf{y}}_1^i))\right] \\ &= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}} \left[\frac{\epsilon^2}{2}\left\|\boldsymbol{\phi}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^i)} - \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^i)}\right\|_2^2\right] \text{ (Using lemma 2.7)} \\ &= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}} \left[\frac{\epsilon^2}{2}\left\|\sum_{j=1}^{i}\left(\prod_{k=j}^{i}\mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}}\right)\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)} - \sum_{j=1}^{i}\mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_F^2\right]\end{aligned}$$

Using lemma 3.1 and assuming $\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)}, \cdots, \boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$ are independent and spherical symmetric

$$= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad \sum_{j=1}^{i}\left[\frac{\epsilon^2}{2|\mathcal{X}|}\left\|\left(\prod_{k=j}^{i}\mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right)\right\|_F^2 \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}}\left[\left\|\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_2^2\right]\right]$$

Since each objective have a unique $\mathbf{H}_j$ and $\mathbf{G}_j$ matrix

$$\begin{aligned}&= \min_{\mathbf{F}} \quad \sum_{j=1}^{i}\min_{\mathbf{H}_j,\mathbf{G}_j}\left[\frac{\epsilon^2}{2|\mathcal{X}|}\left\|\left(\prod_{k=j}^{i}\mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right)\right\|_F^2 \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}}\left[\left\|\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_2^2\right]\right] \\ &= \frac{\alpha\epsilon^2}{2|\mathcal{X}|}\min_{\mathbf{F}} \quad \sum_{j=1}^{i}\min_{\mathbf{H}_j,\mathbf{G}_j}\left[\left\|\left(\prod_{k=j}^{i}\mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right)\right\|_F^2\right] \;;\; \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}}\left[\left\|\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_2^2\right] = \alpha\end{aligned}$$

To solve this optimization problem we will use the Eckart-Young-Mirsky theorem

Theorem 3.2. *(*Eckart-Young-Mirsky Theorem *[47]) Let* $\mathbf{B} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ *have the SVD* $\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ *and we define* $\mathbf{B}^{(r)} = \mathbf{U}^{(r)}\mathbf{S}^{(r)}(\mathbf{V}^{(r)})^T$ *as rank* r *of* $\mathbf{B}$ *associated with the* r *largest singular values and their corresponding singular vectors.* $\mathbf{B}^{(r)}$ *is the optimal solution to the following optimization problem:*

$$\begin{aligned}&\min \;\; ||\mathbf{B} - \hat{\mathbf{B}}|| \\ &\text{s.t.} \;\; \text{rank}(\hat{\mathbf{B}}) \leq r \\ &\implies \hat{\mathbf{B}} = \mathbf{B}^{(r)} = \mathbf{U}^{(r)}\mathbf{S}^{(r)}(\mathbf{V}^{(r)})^T\end{aligned}$$

By using theorem 3.2, the optimal $\mathbf{G}_j, \mathbf{H}_j, \mathbf{F}$ matrices are

$$\mathbf{B}_j = \prod_{k=j}^{i} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} = \mathbf{U}_j\mathbf{S}_j\mathbf{V}_j^T \implies \mathbf{G}_j = \mathbf{U}_j^{(r)} \; ; \; \mathbf{H}_j = \mathbf{S}_j^{(r)}(\mathbf{V}_j^{(r)})^T \; ; \; \mathbf{F} = \mathbf{I}$$

- at node x_{i+1}, we will calculate the posterior distribution

$$\hat{p}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}\left(x_{i+1}|\check{y}_1^i\right) = p_{\mathsf{x}_{i+1}}(x_{i+1}) + \epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}\hat{m}_{\mathsf{x}_i\to\mathsf{x}_{i+1}}(x_{i+1})$$

Based on the optimal matrices, the error due to the constrain in the message is

$$\sum_{j=1}^{i} \left\|\mathbf{B}_j - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right\|_F^2 = \sum_{j=1}^{i}\sum_{t=r+1}^{|\mathcal{X}|} s_t^2(\mathbf{B}_j)$$

In the case of having $\mathbf{B}_j$ as rank r, the error associated with the constrained message will be zero.

3.1.2 Fixed Divergence Transfer Matrix

In this section $\mathbf{B} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ is known exactly and is fixed. The messages are constrained $\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i} \in \mathbb{R}^r$, where $r < |\mathcal{X}|$.

$$\mathbf{B}_{x_i,x_{i+1}} = \mathbf{B} \; ; \; \forall i$$

- at node y_i, pick a matrix $\mathbf{H}_i \in \mathbb{R}^{r\times|\mathcal{X}|}$, to reduce the size of the message ($\mathbf{m}_{\mathsf{y}_i\to\mathsf{x}_i}$) to r-dimensional message ($\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i}$) since there is a constrained link between y_i and x_i.

$$\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i} = \mathbf{H}_i\mathbf{m}_{\mathsf{y}_i\to\mathsf{x}_i} = \mathbf{H}_i\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$$

- at node x_i, pick a matrix $\mathbf{G}_i \in \mathbb{R}^{|\mathcal{X}|\times r}$ and $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ to calculate the forward message

$$\begin{aligned}
\hat{\mathbf{m}}_{\mathsf{x}_i\to\mathsf{x}_{i+1}} &= \mathbf{G}_i\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i} + \mathbf{F}\hat{\mathbf{m}}_{\mathsf{x}_{i-1}\to\mathsf{x}_i} \\
&= \mathbf{G}_i\mathbf{H}_i\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)} + \mathbf{F}\hat{\boldsymbol{\phi}}^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})} \\
&= \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)} + \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^{i-1})} \\
&= \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^{i})}
\end{aligned}$$

- Optimization:

$$\begin{aligned}
&\min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}} \left[D(\mathbf{p}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(.|\check{\mathsf{y}}_1^i)||\hat{\mathbf{p}}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(.|\check{\mathsf{y}}_1^i))\right] \\
&= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}} \left[\frac{\epsilon^2}{2}\left\|\boldsymbol{\phi}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^i)} - \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^i)}\right\|_2^2\right] \text{ (Using lemma 2.7)} \\
&= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}} \left[\frac{\epsilon^2}{2}\left\|\sum_{j=1}^{i}\mathbf{B}^{i+1-j}\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)} - \sum_{j=1}^{i}\mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_F^2\right]
\end{aligned}$$

Using lemma 3.1 and assuming $\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$ are independent and spherical symmetric

$$= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad \sum_{j=1}^{i}\left[\frac{\epsilon^2}{2|\mathcal{X}|}\left\|\left(\mathbf{B}^{i+1-j} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right)\right\|_F^2 \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}}\left[\left\|\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_2^2\right]\right]$$

Since each objective have a unique $\mathbf{H}_j$ and $\mathbf{G}_j$ matrix

$$\begin{aligned}
&= \min_{\mathbf{F}} \quad \sum_{j=1}^{i}\min_{\mathbf{H}_j,\mathbf{G}_j}\left[\frac{\epsilon^2}{2|\mathcal{X}|}\left\|\left(\mathbf{B}^{i+1-j} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right)\right\|_F^2 \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}}\left[\left\|\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_2^2\right]\right] \\
&= \frac{\alpha\epsilon^2}{2|\mathcal{X}|}\min_{\mathbf{F}} \quad \sum_{j=1}^{i}\min_{\mathbf{H}_j,\mathbf{G}_j}\left[\left\|\left(\mathbf{B}^{i+1-j} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right)\right\|_F^2\right] \ ; \ \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}}\left[\left\|\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_2^2\right] = \alpha
\end{aligned}$$

By using theorem 3.2, the optimal $\mathbf{G}_j, \mathbf{H}_j, \mathbf{F}$ matrices

$$\mathbf{B}_j = \mathbf{B}^{i+1-j} = \mathbf{U}_j\mathbf{S}_j\mathbf{V}_j^T \implies \mathbf{G}_j = \mathbf{U}_j^{(r)} \ ; \ \mathbf{H}_j = \mathbf{S}_j^{(r)}(\mathbf{V}_j^{(r)})^T \ ; \ \mathbf{F} = \mathbf{I}$$

- at node x_{i+1}, we will calculate the posterior distribution

$$\hat{p}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}\left(x_{i+1}|\check{y}_1^i\right) = p_{\mathsf{x}_{i+1}(x_{i+1})} + \epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}\hat{m}_{\mathsf{x}_i\to\mathsf{x}_{i+1}}(x_{i+1})$$

To derive a sharp upper bound of the error we will define the following theorem

Theorem 3.3. *[48]: Let* $\mathbf{A}$ *and* $\mathbf{C}$ *be* $m \times n$ *matrices with singular values* $\{s_i(\mathbf{A})\}_{i=1}^q$ *and* $\{s_i(\mathbf{C})\}_{i=1}^q$, $q = \min\{m, n\}$ *and left singular vectors* $\{\mathbf{u}_i(\mathbf{C})\}_{i=1}^n$ *and right singular vectors* $\{\mathbf{v}_i(\mathbf{A})\}_{i=1}^n$*. If for some* r *and* s, $0 \leq r, s \leq q-1$,

$$dim(\langle u_1, \cdots, u_r\rangle \cap \langle v_1, \cdots, v_s\rangle) \geq k, \ k \geq 0$$

then

$$s_{i+j-k-1}(\mathbf{AC}) \leq s_i(\mathbf{A})s_j(\mathbf{C}), \quad r+1 \leq i \leq q, \quad s+1 \leq j \leq q, \quad i+j-k-1 \leq q$$

Since $\mathbf{B}$ is fixed, the first left singular vector $u_1(\mathbf{B})$ and first right singular vector $v_1(\mathbf{B})$ are equal and so $k = 1$. Therefore, based on the optimal matrices the error due to the constrain in the message is

$$\begin{aligned}
\sum_{j=1}^{i} ||\mathbf{B}_j - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j||_F^2 &= \sum_{j=1}^{i}\sum_{t=r+1}^{|\mathcal{X}|} s_t^2(\mathbf{B}_j) \\
&= \sum_{j=1}^{i}\sum_{t=r+1}^{|\mathcal{X}|} s_t^2(\mathbf{B}^{i+1-j}) \\
&\leq \sum_{j=1}^{i}\sum_{t=r+1}^{|\mathcal{X}|} s_2^{2(i-j)}(\mathbf{B})s_t^2(\mathbf{B}) \ \ \text{Using Theorem 3.3}
\end{aligned}$$

Since $s_2(\mathbf{B}) < 1$, this means that the error associated with older observation will decay at an exponential rate faster than $s_2^2(\mathbf{B})$.

3.1.3 Computation and Communication Complexity

The performance of the algorithm is expected to be sub-optimal due to the constrained messages and the insufficient statistics. However, these constraints have the advantage of reducing computational and communication complexity, which is particularly useful in modern applications where high-dimensional data is prevalent.

Table 3.2 provides a comparison of computational and communication complexity for the constrained and unconstrained message passing algorithms. As shown in the table, the computational complexity is reduced from $O(|\mathcal{X}|^2)$ to $O(r|\mathcal{X}|)$, where $r < |\mathcal{X}|$. This reduction in complexity is significant, particularly for models with large states, such as language or vision models.

In the preceding algorithm, the DTM is utilized within the context of single index-specific posterior inference. As a defining characteristic of the DTM model, the largest singular value is 1. This implies the necessity to incorporate all the previous observations (n), yielding a total computational complexity of $O(nr|\mathcal{X}|)$.

Conversely, as described in the prior chapter, there exists an alternative approach through the use of CDM. In the CDM, the largest singular value is less than or equal to 1. This property substantiates the possibility of dismissing the observations from earlier stages as the singular values associated with the CDM product matrix will diminish towards an extremely small magnitude.

To elucidate further, in contrast to the DTM inherent requirement to consider all preceding data points (as a direct result of its leading singular value of 1), the CDM leading singular value being less than or equal to 1 implies that it does not attribute significant weight to the older observations. This particular attribute of the CDM can be of substantial advantage in scenarios where computational efficiency is crucial.

$$||\hat{\mathbf{B}}^{n_1}||_F^2 \approx 0 \;;\; ||\prod_{i=1}^{n_1} \hat{\mathbf{B}}_{\mathrm{x_i},\mathrm{x_{i+1}}}||_F^2 \approx 0$$

Furthermore, the communication complexity for $\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i}$is reduced from $O(|\mathcal{X}|)$ to $O(r)$, which is particularly useful in the case of observation nodes with limited communication capabilities. These advantages come at the cost of a sub-optimal algorithm performance when compared to the case of unconstrained messages.

Despite the sub-optimality of the algorithm, it remains a practical and useful solution for high-dimensional data problems, particularly when computational and communication resources are limited. This scenario highlights the importance of considering the trade-off between algorithm performance and resource utilization in modern data-driven applications.

		Constrained Message Passing	Unconstrained Message Passing
Computational Complexity	$\mathbf{m}_{\mathsf{x}_i \to \mathsf{x}_{i+1}}$	$O\left(r\lvert\mathcal{X}\rvert\right)$	$O\left(\lvert\mathcal{X}\rvert^2\right)$
	Total	$O\left(n_2 r\lvert\mathcal{X}\rvert\right)$	$O\left(n\lvert\mathcal{X}\rvert^2\right)$
Communication Complexity	$\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i}$	$O\left(r\right)$	$O\left(\lvert\mathcal{X}\rvert\right)$
	$\mathbf{m}_{\mathsf{x}_i \to \mathsf{x}_{i+1}}$	$O\left(\lvert\mathcal{X}\rvert\right)$	$O\left(\lvert\mathcal{X}\rvert\right)$

Table 3.1: Computational and communication complexity comparison between the constrained and unconstrained message passing algorithm in single index-specific posterior inference. n is the number of observation used for the posterior predication and $n_2 = \min\{n, n_1\}$

3.1.4 Numerical Results

This section aims to explore the performance of the constrained information exchange message passing algorithm in single index-specific posterior inference, using synthetic dataset. The focus will be on the case $D(\mathbf{p}_{\mathsf{x}_3|\mathsf{y}_1^2}||\hat{\mathbf{p}}_{\mathsf{x}_3|\mathsf{y}_1^2})$ as illustrated in Fig. 3-2.

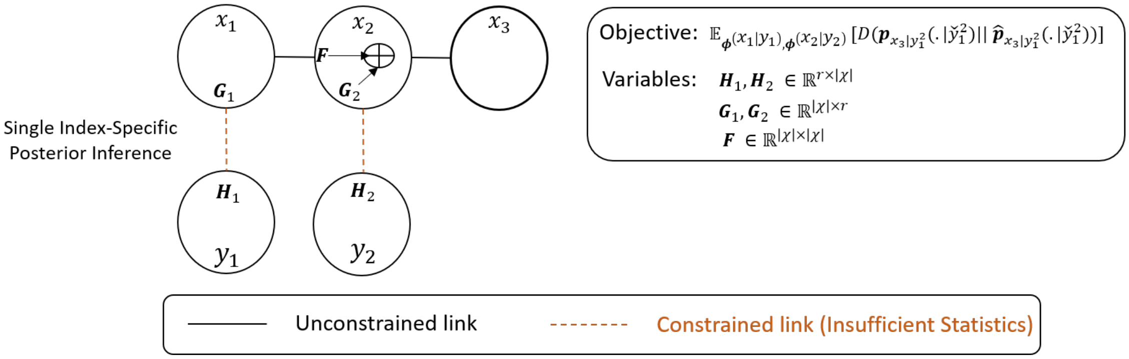

x_1
x_2
x_3
$\boldsymbol{G}_1$
$\boldsymbol{F}$
$\boldsymbol{G}_2$
Single Index-Specific Posterior Inference
$\boldsymbol{H}_1$
y_1
$\boldsymbol{H}_2$
y_2
Objective: $\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2)}[D(\boldsymbol{p}_{x_3|y_1^2}(.|\check{y}_1^2)||\hat{\boldsymbol{p}}_{x_3|y_1^2}(.|\check{y}_1^2))]$
Variables: $\boldsymbol{H}_1, \boldsymbol{H}_2 \in \mathbb{R}^{r\times|\chi|}$
$\boldsymbol{G}_1, \boldsymbol{G}_2 \in \mathbb{R}^{|\chi|\times r}$
$\boldsymbol{F} \in \mathbb{R}^{|\chi|\times|\chi|}$
Unconstrained link
Constrained link (Insufficient Statistics)

Figure 3-2: Constrained Information Exchange Message-Passing Algorithm in Single Index-Specific Posterior Inference

3.1.4.1 Synthetic dataset

The synthetic dataset is strategically designed to possess a large state space ($|\mathcal{X}| = 200$) and observation space ($|\mathcal{Y}| = 20k$), underscoring the critical role of the constrained information exchange message passing algorithm in managing high dimensional datasets. The hidden node x_i and the observation node y_i in this synthetic dataset is weakly correlated.

Dependency of Error on Link Size: Fig. 3-3 illustrates the dependency between the error and the link size. As observed, there is a consistent decrease in error with an increase in the link size. This trend is attributed to the larger amount of information exchanged as the link size expands. Ultimately, the error drops to zero when the link size equals the rank of the DTM matrix.

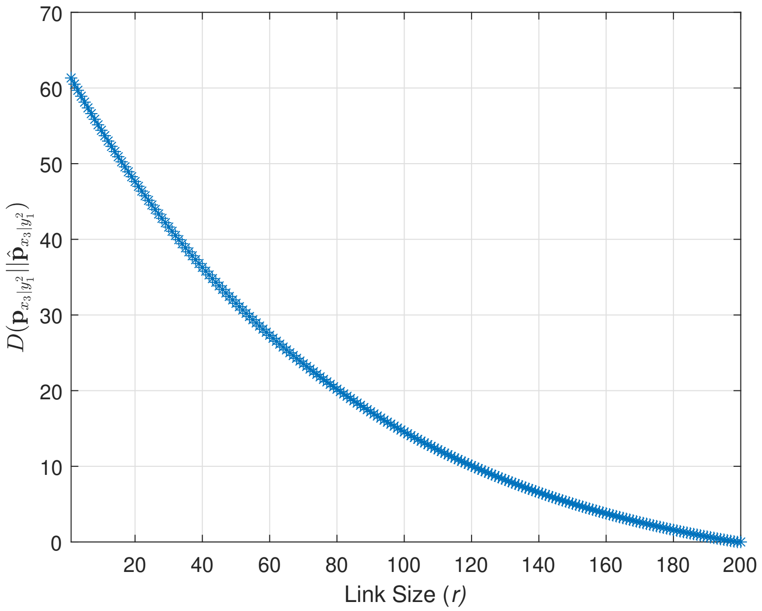


Figure 3-3: Single Index-Specific Posterior Inference error as a function of link size

Determining Link Sizes: Guided by a Predetermined Error Threshold: Fig. 3-4 examines a scenario in which the objective is to limit the error below a preset threshold. This situation calls for a strategic choice of link size that aligns with this error constraint. As illustrated in Figure 3-4, maintaining the error below the 5 threshold demands a minimum link size of 151 ($r \geq 151$). Therefore, while maintaining the error threshold the reduction in the size of the link is by almost **25%**. This emphasizes the significance of constrained information message passing algorithm to deal with high-dimensional data.

Link Size Selection: Shaped by Error Threshold and Size Constraints: Fig. 3-5 illustrates the scenarios where the task is to maintain the error below a particular threshold, whilst staying within specific link size constraints. Such circumstances call for an intricate balance between these two defining parameters. As depicted in Fig. 3-5, if the target is to cap the error at less than 5, with the link size not surpassing 155, the suitable range for link size ($151 \leq r \leq 155$). This underscores the significance of precise link size selection, which is shaped by both the desired error threshold and the maximum size limitations.

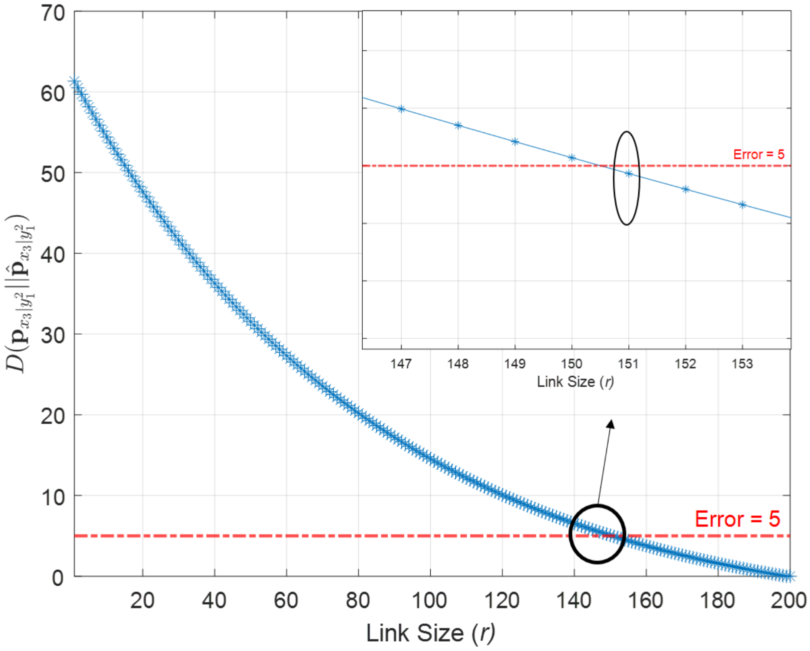


Figure 3-4: Link Sizes Selection Based on Desired Error Threshold in the Single Index-Specific Posterior Inference

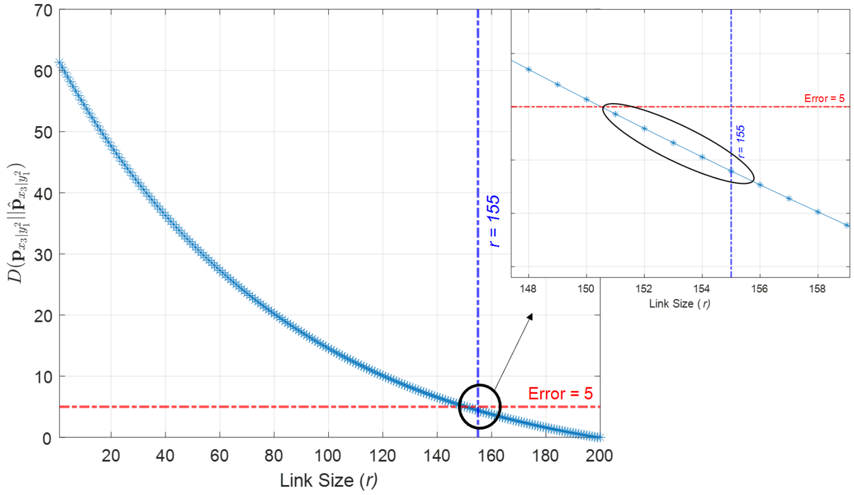


Figure 3-5: Link Size Selection Based on Desired Error Threshold and Maximum Size Constraint in the Single Index-Specific Posterior Inference

Error Threshold Impact on Minimum Link Size: Fig. 3-6 illustrate the tradeoff between the error threshold and the corresponding minimum link size. As observed, there is a consistent decrease in the minimum link size with an increase in the associated error. This trend is attributed to the smaller amount of information exchanged as the link size decreases. Fig. 3-6 emphasizes the balance between optimizing resource utilization and achieving the necessary precision. As the size of the link decreases, the system becomes more efficient in terms of resource usage. However, it also necessitates a higher error, potentially affecting the overall performance and accuracy of the system. The figure presents a valuable perspective on system design considerations in the constrained information exchange message passing algorithm in high dimensional data.

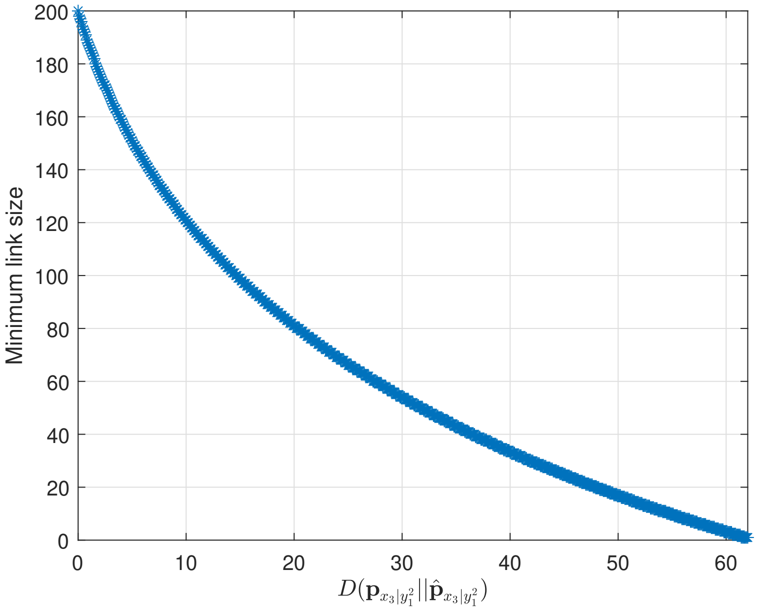


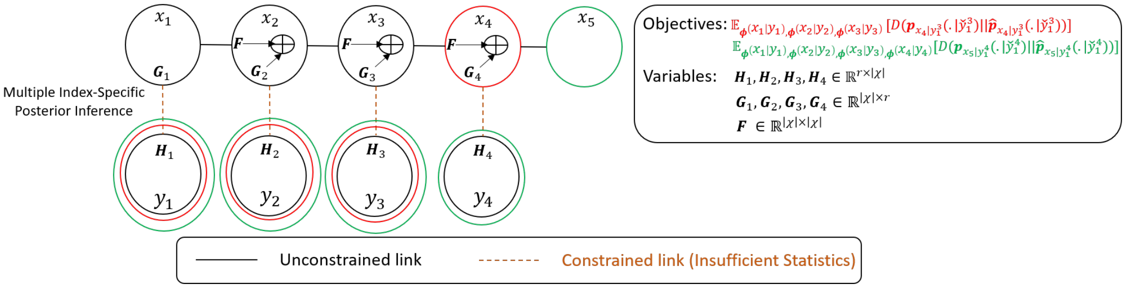
Figure 3-6: Error Threshold Impact on Minimum Link Size

3.2 Multiple Index-Specific Posterior Inference

This scenario involves estimating the posterior distribution at a multiple specific indices, where a multi-objective optimization problem is solved to determine the optimal matrices for compressing messages $\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i}$ (as show in Fig. 3-7). This scenario unveils the inherent tradeoff nature of the problem arising from the inability to transmit sufficient statistics. Two cases will be studied in this scenario:

- Two index-specific posterior inference with varying DTM matrix.

- Multiple index-specific posterior inference with fixed DTM matrix



x_1
x_2
x_3
x_4
x_5
$\boldsymbol{F}$
$\boldsymbol{G}_1$
$\boldsymbol{G}_2$
$\boldsymbol{G}_3$
$\boldsymbol{G}_4$
Multiple Index-Specific Posterior Inference
$\boldsymbol{H}_1$
$\boldsymbol{H}_2$
$\boldsymbol{H}_3$
$\boldsymbol{H}_4$
y_1
y_2
y_3
y_4
Objectives: $\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2),\boldsymbol{\phi}(x_3|y_3)}\left[D(\boldsymbol{p}_{x_4|y_1^3}(.|\check{y}_1^3)||\widehat{\boldsymbol{p}}_{x_4|y_1^3}(.|\check{y}_1^3))\right]$
$\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2),\boldsymbol{\phi}(x_3|y_3),\boldsymbol{\phi}(x_4|y_4)}\left[D(\boldsymbol{p}_{x_5|y_1^4}(.|\check{y}_1^4)||\widehat{\boldsymbol{p}}_{x_5|y_1^4}(.|\check{y}_1^4))\right]$
Variables: $\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3,\boldsymbol{H}_4 \in \mathbb{R}^{r\times|\chi|}$
$\boldsymbol{G}_1,\boldsymbol{G}_2,\boldsymbol{G}_3,\boldsymbol{G}_4 \in \mathbb{R}^{|\chi|\times r}$
$\boldsymbol{F} \in \mathbb{R}^{|\chi|\times|\chi|}$
Unconstrained link
Constrained link (Insufficient Statistics)

Figure 3-7: Constrained Information Exchange Message-Passing Algorithm in Two Index-Specific Posterior Inference

3.2.1 Two Index-Specific Posterior Inference with Varying DTM Matrix

In this section $\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ is known exactly and the messages are constrained $\hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i} \in \mathbb{R}^r$, where $r < |\mathcal{X}|$.

- at node y_i, pick a matrix $\mathbf{H}_i \in \mathbb{R}^{r\times|\mathcal{X}|}$, to reduce the size of the message $(\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i})$ to r-dimensional message $(\hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i})$ since there is a constrained link between y_i and x_i.

$$\hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i} = \mathbf{H}_i \mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i} = \mathbf{H}_i \boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$$

- at node x_i, pick a matrix $\mathbf{G}_i \in \mathbb{R}^{|\mathcal{X}|\times r}$ and $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ to calculate the forward message

$$\begin{aligned}
\hat{\mathbf{m}}_{\mathsf{x}_i\to\mathsf{x}_{i+1}} &= \mathbf{G}_i\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i} + \mathbf{F}\hat{\mathbf{m}}_{\mathsf{x}_{i-1}\to\mathsf{x}_i} \\
&= \mathbf{G}_i\mathbf{H}_i\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)} + \mathbf{F}\hat{\boldsymbol{\phi}}^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})} \\
&= \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)} + \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^{i-1})} \\
&= \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^{i})}
\end{aligned}$$

- First objective optimization:

$$\begin{aligned}
&\min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}}\left[D(\mathbf{p}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(.|\check{\mathsf{y}}_1^i)||\hat{\mathbf{p}}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(.|\check{\mathsf{y}}_1^i))\right] \\
&= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}}\left[\frac{\epsilon^2}{2}\left\|\boldsymbol{\phi}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^i)} - \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^i)}\right\|_2^2\right] \text{ (Using lemma 2.7)} \\
&= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}}\left[\frac{\epsilon^2}{2}\left\|\sum_{j=1}^{i}\left(\prod_{k=j}^{i}\mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}}\right)\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)} - \sum_{j=1}^{i}\mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_F^2\right]
\end{aligned}$$

Using lemma 3.1 and assuming $\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$ are independent and spherical symmetric

$$\begin{aligned}
&= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \sum_{j=1}^{i}\left[\frac{\epsilon^2}{2|\mathcal{X}|}\left\|\left(\prod_{k=j}^{i}\mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right)\right\|_F^2 \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}}\left[\left\|\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_2^2\right]\right] \\
&= \frac{\alpha\epsilon^2}{2|\mathcal{X}|}\min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \sum_{j=1}^{i}\left[\left\|\left(\prod_{k=j}^{i}\mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right)\right\|_F^2\right]
\end{aligned}$$

- Second objective optimization

$$\begin{aligned}
&\min_{\mathbf{H}_1,\cdots,\mathbf{H}_{i-1},\mathbf{G}_1,\cdots,\mathbf{G}_{i-1},\mathbf{F}} \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_{i-1}|\mathsf{y}_{i-1})}}\left[D(\mathbf{p}_{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(.|\check{\mathsf{y}}_1^{i-1})||\hat{\mathbf{p}}_{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(.|\check{\mathsf{y}}_1^{i-1}))\right] \\
&= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_{i-1},\mathbf{G}_1,\cdots,\mathbf{G}_{i-1},\mathbf{F}} \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_{i-1}|\mathsf{y}_{i-1})}}\left[\frac{\epsilon^2}{2}\left\|\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})} - \hat{\boldsymbol{\phi}}^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})}\right\|_2^2\right] \text{ (Using lemma 2.7)} \\
&= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_{i-1},\mathbf{G}_1,\cdots,\mathbf{G}_{i-1},\mathbf{F}} \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_{i-1}|\mathsf{y}_{i-1})}}\left[\frac{\epsilon^2}{2}\left\|\sum_{j=1}^{i-1}\left(\prod_{k=j}^{i-1}\mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}}\right)\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)} - \sum_{j=1}^{i-1}\mathbf{F}^{i-1-j}\mathbf{G}_j\mathbf{H}_j\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}\right\|_F^2\right]
\end{aligned}$$

Using lemma 3.1 and assuming $\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)}, \cdots, \boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$ are independent and spherical symmetric

$$= \min_{\mathbf{H}_1,\cdots,\mathbf{H}_{i-1},\mathbf{G}_1,\cdots,\mathbf{G}_{i-1},\mathbf{F}} \quad \sum_{j=1}^{i-1} \left[\frac{\epsilon^2}{2|\mathcal{X}|} \left\| \left(\prod_{k=j}^{i-1} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-1-j}\mathbf{G}_j\mathbf{H}_j \right) \right\|_F^2 \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}} \left[\left\| \boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)} \right\|_2^2 \right] \right]$$

$$= \frac{\alpha\epsilon^2}{2|\mathcal{X}|} \min_{\mathbf{H}_1,\cdots,\mathbf{H}_{i-1},\mathbf{G}_1,\cdots,\mathbf{G}_{i-1},\mathbf{F}} \quad \sum_{j=1}^{i-1} \left[\left\| \left(\prod_{k=j}^{i-1} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-1-j}\mathbf{G}_j\mathbf{H}_j \right) \right\|_F^2 \right]$$

- There are several way to solve this multiobjective optimization. We will write these objective functions as a **weighted sum** and then we will solve the weighted sum for different values of γ [49].

$$\begin{aligned}
\min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad & \gamma \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_{i-1}|\mathsf{y}_{i-1})}} \left[D(\mathbf{p}_{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(.|\check{\mathsf{y}}_1^{i-1}) || \hat{\mathbf{p}}_{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(.|\check{\mathsf{y}}_1^{i-1})) \right] \\
& + (1-\gamma) \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}} \left[D(\mathbf{p}_{\mathsf{x}_{i+1}|\mathsf{y}_1^{i}}(.|\check{\mathsf{y}}_1^{i}) || \hat{\mathbf{p}}_{\mathsf{x}_{i+1}|\mathsf{y}_1^{i}}(.|\check{\mathsf{y}}_1^{i})) \right] \\
= \frac{\alpha\epsilon^2}{2|\mathcal{X}|} \min_{\mathbf{H}_1,\cdots,\mathbf{H}_{i-1},\mathbf{G}_1,\cdots,\mathbf{G}_{i-1},\mathbf{F}} \quad & \gamma \sum_{j=1}^{i-1} \left[\left\| \left(\prod_{k=j}^{i-1} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-1-j}\mathbf{G}_j\mathbf{H}_j \right) \right\|_F^2 \right] \\
& + (1-\gamma) \sum_{j=1}^{i-1} \left[\left\| \left(\prod_{k=j}^{i} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j \right) \right\|_F^2 \right] \\
& + \frac{\alpha\epsilon^2}{2|\mathcal{X}|} \min_{\mathbf{H}_i,\mathbf{G}_i} ||\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}} - \mathbf{G}_i\mathbf{H}_i||_F^2 \\
= \frac{\alpha\epsilon^2}{2|\mathcal{X}|} \min_{\mathbf{H}_1,\cdots,\mathbf{H}_{i-1},\mathbf{G}_1,\cdots,\mathbf{G}_{i-1},\mathbf{F}} \quad & \sum_{j=1}^{i-1} \left\| \begin{bmatrix} \sqrt{\gamma}(\left(\prod_{k=j}^{i-1} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-1-j}\mathbf{G}_j\mathbf{H}_j\right)) \\ \sqrt{(1-\gamma)} \left(\prod_{k=j}^{i} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right) \end{bmatrix} \right\|_F^2 \\
& + \frac{\alpha\epsilon^2}{2|\mathcal{X}|} \min_{\mathbf{H}_i,\mathbf{G}_i} ||\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}} - \mathbf{G}_i\mathbf{H}_i||_F^2
\end{aligned}$$

The following optimization have two parts, the minimization with $\mathbf{G}_i$ and $\mathbf{H}_i$ can be solved using theorem 3.2. The other minimization is bit more complex and can be solved using the following theorem

Theorem 3.4. *Let* $\mathbf{B}_{x_j,x_{j+1}} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$, $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$,$\mathbf{G}_j \in \mathbb{R}^{|\mathcal{X}|\times r}$, *and* $\mathbf{H}_j \in \mathbb{R}^{r\times|\mathcal{X}|}$ *where* $j \in \{1, \cdots, i\}$*. The SVD of*

$$\left\|\begin{bmatrix} \sqrt{\gamma}(\left(\prod_{k=j}^{i-1} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}}\right)) \\ \sqrt{(1-\gamma)}\left(\prod_{k=j}^{i} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}}\right) \end{bmatrix}\right\|_F^2 = \begin{bmatrix} \underbrace{\mathbf{U}_{j,1}}_{|\mathcal{X}|\times|\mathcal{X}|} & \\ & \underbrace{\mathbf{G}}_{2|\mathcal{X}|\times|\mathcal{X}|} \\ \underbrace{\mathbf{U}_{j,2}}_{|\mathcal{X}|\times|\mathcal{X}|} & \end{bmatrix} \begin{bmatrix} \underbrace{\overline{\mathbf{S}}_j}_{|\mathcal{X}|\times|\mathcal{X}|} \\ \mathbf{0}_{|\mathcal{X}|\times|\mathcal{X}|} \end{bmatrix} \underbrace{\mathbf{V}_j^T}_{|\mathcal{X}|\times|\mathcal{X}|}$$

Therefore, the optimal matrices for

$$\left\|\begin{bmatrix} \sqrt{\gamma}(\left(\prod_{k=j}^{i-1} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-1-j}\mathbf{G}_j\mathbf{H}_j\right)) \\ \sqrt{(1-\gamma)}\left(\prod_{k=j}^{i} \mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j\right) \end{bmatrix}\right\|_F^2$$

are

$$\begin{aligned} \mathbf{H}_j &= \overline{\mathbf{S}}_j^{(r)}\mathbf{V}_j^{(r)T} \\ \mathbf{G}_j &= \frac{1}{\sqrt{\gamma}}(\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}}^{-1})^{i-1-j}\mathbf{U}_{j,1}\begin{bmatrix} \mathbf{I}_{r\times r} \\ 0_{|\mathcal{X}|-r\times r} \end{bmatrix} \\ &= \frac{1}{\sqrt{\gamma}}(\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}})^{j+1-i}\mathbf{U}_{j,1}\begin{bmatrix} \mathbf{I}_{r\times r} \\ 0_{|\mathcal{X}|-r\times r} \end{bmatrix} \\ \mathbf{F} &= \mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}} \end{aligned}$$

where $\overline{\mathbf{S}}_j^{(r)}$ *is* $r \times r$ *matrix with* r *largest singular values,* $\mathbf{V}_j^{(r)}$ *is* $|\mathcal{X}| \times r$ *matrix with the corresponding right singular vectors.*

Proof. The eigenvalue decomposition of each of the concatenated matrices

$$\begin{aligned} \sqrt{\gamma}\prod_{k=j}^{i-1} \mathbf{B}_{x_k,x_{k+1}} &= \sqrt{\gamma}\prod_{k=j}^{i-1} \mathbf{Q}_k\mathbf{\Delta}_k\mathbf{Q}_k^{-1} \\ \sqrt{1-\gamma}\prod_{k=j}^{i} \mathbf{B}_{x_k,x_{k+1}} &= \sqrt{1-\gamma}\prod_{k=j}^{i} \mathbf{Q}_k\mathbf{\Delta}_k\mathbf{Q}_k^{-1} \end{aligned}$$

The singular value decomposition of these two concatenated matrices

$$\sqrt{\gamma}\prod_{k=j}^{i-1}\mathbf{B}_{x_k,x_{k+1}} = \mathbf{U}_{j,1}\overline{\mathbf{S}}_j\mathbf{V}_j^T$$
$$\sqrt{1-\gamma}\prod_{k=j}^{i}\mathbf{B}_{x_k,x_{k+1}} = \mathbf{U}_{j,2}\overline{\mathbf{S}}_j\mathbf{V}_j^T$$

We will write the left singular vector in terms of the eigenvalue decomposition, $\overline{\mathbf{S}}_j$, and $\mathbf{V}_j^T$

$$\mathbf{U}_{j,1} = \sqrt{\gamma}\prod_{k=j}^{i-1}\mathbf{Q}_k\boldsymbol{\Delta}_k\mathbf{Q}_k^{-1}\mathbf{V}_j\overline{\mathbf{S}}_j^{-1}$$
$$\mathbf{U}_{j,2} = \sqrt{1-\gamma}\prod_{k=j}^{i}\mathbf{Q}_k\boldsymbol{\Delta}_k\mathbf{Q}_k^{-1}\mathbf{V}_j\overline{\mathbf{S}}_j^{-1}$$
$$\mathbf{U}_{j,2}(\mathbf{U}_{j,1})^{-1} = \sqrt{\frac{1-\gamma}{\gamma}}\mathbf{Q}_i\boldsymbol{\Delta}_i\mathbf{Q}_i^{-1} = \sqrt{\frac{1-\gamma}{\gamma}}\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}}$$

Therefore, we realize that the left singular vectors are related in terms of a scaling factor $\sqrt{\frac{1-\gamma}{\gamma}}$ and the matrix $\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}}$. Since, the difference between the two concatenated matrices is just an $\mathbf{F}$, then

$$\mathbf{F} = \mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}}$$

The singular value decomposition we showed above of the concatenated matrices is the optimal low rank approximation. The only issue is that the $\mathbf{F}$ might have an effect and so we will incorporate the inverse of the $\mathbf{F}$ in the $\mathbf{G}_j$ matrix as the following

$$\mathbf{G}_j = \frac{1}{\sqrt{\gamma}}(\mathbf{B}_{\mathsf{x}_i,\mathsf{x}_{i+1}})^{j+1-i}\mathbf{U}_{j,1}\begin{bmatrix}\mathbf{I}_{r\times r}\\ 0_{|\mathcal{X}|-r\times r}\end{bmatrix}$$

$$\mathbf{H}_j = \overline{\mathbf{S}}_j^{(r)} \mathbf{V}_j^{(r)T}$$

□

In this scenario, a proof is presented that utilizes concepts from singular value decomposition and eigenvalue decomposition to establish a relationship between them in the case of a concatenated matrix. The relationship between these two concepts is fundamental in understanding the underlying structure of matrices, which has a significant impact on low rank approximation of matrices.

3.2.2 Multiple Index-Specific Posterior Inference with Fixed DTM

In this section $\mathbf{B} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ is known exactly and is fixed. The messages are constrained $\hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i} \in \mathbb{R}^r$, where $r < |\mathcal{X}|$.

- at node y_i, pick a matrix $\mathbf{H}_i \in \mathbb{R}^{r \times |\mathcal{X}|}$, to reduce the size of the message $(\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i})$ to r-dimensional message $(\hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i})$ since there is a constrained link between y_i and x_i.

$$\hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i} = \mathbf{H}_i \mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i} = \mathbf{H}_i \boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$$

- at node x_i, pick a matrix $\mathbf{G}_i \in \mathbb{R}^{|\mathcal{X}| \times r}$ and $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ to calculate the forward message

$$\begin{aligned} \hat{\mathbf{m}}_{\mathsf{x}_i \to \mathsf{x}_{i+1}} &= \mathbf{G}_i \hat{\mathbf{m}}_{\mathsf{y}_i \to \mathsf{x}_i} + \mathbf{F} \hat{\mathbf{m}}_{\mathsf{x}_{i-1} \to \mathsf{x}_i} \\ &= \mathbf{G}_i \mathbf{H}_i \boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)} + \mathbf{F} \hat{\boldsymbol{\phi}}^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})} \\ &= \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)} + \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^{i-1})} \\ &= \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^{i})} \end{aligned}$$

- Optimization objectives

$$\min_{\mathbf{H}_1, \cdots, \mathbf{H}_i, \mathbf{G}_1, \cdots, \mathbf{G}_i, \mathbf{F}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)}, \cdots, \boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}} \left[D(\mathbf{p}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(.|\check{\mathsf{y}}_1^i) || \hat{\mathbf{p}}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(.|\check{\mathsf{y}}_1^i)) \right]$$

$$\min_{\mathbf{H}_1,\cdots,\mathbf{H}_{i-1},\mathbf{G}_1,\cdots,\mathbf{G}_{i-1},\mathbf{F}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_{i-1}|\mathsf{y}_{i-1})}} \left[D(\mathbf{p}_{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(.|\check{\mathsf{y}}_1^{i-1})||\hat{\mathbf{p}}_{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(.|\check{\mathsf{y}}_1^{i-1})) \right]$$

$$\vdots$$

$$\min_{\mathbf{H}_1,\mathbf{G}_1} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)}} \left[D(\mathbf{p}_{\mathsf{x}_2|\mathsf{y}_1}(.|\check{\mathsf{y}}_1)||\hat{\mathbf{p}}_{\mathsf{x}_2|\mathsf{y}_1}(.|\check{\mathsf{y}}_1)) \right]$$

- Using lemma 2.7 and 3.1, and assuming that the information vectors are spherical symmetric and independent

$$\min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad \frac{\alpha\epsilon^2}{2m} \sum_{j=1}^{i} \left\| \mathbf{B}^{i-j+1} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j \right\|_F^2$$

$$\min_{\mathbf{H}_1,\cdots,\mathbf{H}_{i-1},\mathbf{G}_1,\cdots,\mathbf{G}_{i-1},\mathbf{F}} \quad \frac{\alpha\epsilon^2}{2m} \sum_{j=1}^{i-1} \left\| \mathbf{B}^{i-j} - \mathbf{F}^{i-j-1}\mathbf{G}_j\mathbf{H}_j \right\|_F^2$$

$$\vdots$$

$$\min_{\mathbf{H}_1,\mathbf{G}_1} \quad \frac{\alpha\epsilon^2}{2m} \left\| \mathbf{B} - \mathbf{G}_1\mathbf{H}_1 \right\|_F^2$$

- We will write these objective functions as a weighted sum and then we will solve the weighted sum

$$\begin{aligned} \frac{\alpha\epsilon^2}{2m} \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad & \Bigg[\gamma_1 \sum_{j=1}^{i} \left\| \mathbf{B}^{i-j+1} - \mathbf{F}^{i-j}\mathbf{G}_j\mathbf{H}_j \right\|_F^2 + \gamma_2 \sum_{j=1}^{i-1} \left\| \mathbf{B}^{i-j} - \mathbf{F}^{i-j-1}\mathbf{G}_j\mathbf{H}_j \right\|_F^2 \\ & + \cdots + \gamma_i \left\| \mathbf{B} - \mathbf{G}_1\mathbf{H}_1 \right\|_F^2 \Bigg] \\ \text{s.t} \qquad & \sum_i \gamma_i = 1 \end{aligned}$$

- The summation will be rearranged so that objectives with the same matrices $\mathbf{G}_j$ and $\mathbf{H}_j$ will be grouped together.

$$\frac{\alpha\epsilon^2}{2m} \min_{\mathbf{H}_1,\cdots,\mathbf{H}_i,\mathbf{G}_1,\cdots,\mathbf{G}_i,\mathbf{F}} \quad \sum_{j=1}^{i} \sum_{k=1}^{i-j+1} \gamma_k \left\| \mathbf{B}^{i-j-k+2} - \mathbf{F}^{i-j-k+1}\mathbf{G}_j\mathbf{H}_j \right\|_F^2$$

- The objectives with the same matrices $\mathbf{G}_j$ and $\mathbf{H}_j$ will be concatenated together.

Without loss of generality, the concatenated matrix will be shown for $j = 1, 2, i$

$$\left\| \begin{matrix} \sqrt{\gamma_1}(\mathbf{B}^i - \mathbf{F}^{i-1}\mathbf{G}_1\mathbf{H}_1) \\ \sqrt{\gamma_2}(\mathbf{B}^{i-1} - \mathbf{F}^{i-2}\mathbf{G}_1\mathbf{H}_1) \\ \vdots \\ \sqrt{\gamma_i}(\mathbf{B} - \mathbf{G}_1\mathbf{H}_1) \end{matrix} \right\|_F^2 \tag{3.1}$$

$$\left\| \begin{matrix} \sqrt{\gamma_1}(\mathbf{B}^{i-1} - \mathbf{F}^{i-2}\mathbf{G}_2\mathbf{H}_2) \\ \sqrt{\gamma_2}(\mathbf{B}^{i-2} - \mathbf{F}^{i-3}\mathbf{G}_2\mathbf{H}_2) \\ \vdots \\ \sqrt{\gamma_{i-1}}(\mathbf{B} - \mathbf{G}_2\mathbf{H}_2) \end{matrix} \right\|_F^2 \tag{3.2}$$

$$\left\| \sqrt{\gamma_1}(\mathbf{B} - \mathbf{G}_i\mathbf{H}_i) \right\|_F^2 \tag{3.3}$$

where the size of the concatenated matrix depends on the value of j. More precisely, the size of the matrix is $(i + 1 - j)|\mathcal{X}| \times |\mathcal{X}|$.

- The optimization problem can be solved using the following theorem which is very similar to theorem 3.4.

Theorem 3.5. *Let* $\mathbf{B} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$*,* $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$*,*$\mathbf{G}_j \in \mathbb{R}^{|\mathcal{X}|\times r}$*, and* $\mathbf{H}_j \in \mathbb{R}^{r\times|\mathcal{X}|}$ *where* $j \in \{1, \cdots, i\}$*. Without loss of generality, the SVD of the concatenated matrix at* $j = 1$

$$\left\| \begin{matrix} \sqrt{\gamma_1}\mathbf{B}^i \\ \sqrt{\gamma_2}\mathbf{B}^{i-1} \\ \vdots \\ \sqrt{\gamma_i}\mathbf{B} \end{matrix} \right\|_F^2 = \begin{bmatrix} \underbrace{\mathbf{U}_{1,1}}_{|\mathcal{X}|\times|\mathcal{X}|} & \\ \underbrace{\mathbf{U}_{1,2}}_{|\mathcal{X}|\times|\mathcal{X}|} & \underbrace{\mathbf{G}}_{(i)|\mathcal{X}|\times|\mathcal{X}|} \\ \vdots & \\ \underbrace{\mathbf{U}_{1,i}}_{|\mathcal{X}|\times|\mathcal{X}|} & \end{bmatrix} \begin{bmatrix} \underbrace{\overline{\mathbf{S}}_1}_{|\mathcal{X}|\times|\mathcal{X}|} \\ \mathbf{0}_{(i-1)|\mathcal{X}|\times|\mathcal{X}|} \end{bmatrix} \underbrace{\mathbf{V}_1^T}_{|\mathcal{X}|\times|\mathcal{X}|}$$

where $\mathbf{U}_{1,t}$ *and* $\mathbf{V}_1$ *are the left and right singular vectors, and* $\overline{\mathbf{S}}_1$ *are the singular values. The optimal matrices for*

$$\left\| \begin{matrix} \sqrt{\gamma_1}(\mathbf{B}^i - \mathbf{F}^{i-1}\mathbf{G}_1\mathbf{H}_1) \\ \sqrt{\gamma_2}(\mathbf{B}^{i-1} - \mathbf{F}^{i-2}\mathbf{G}_1\mathbf{H}_1) \\ \vdots \\ \sqrt{\gamma_i}(\mathbf{B} - \mathbf{G}_1\mathbf{H}_1) \end{matrix} \right\|_F^2$$

are

$$\mathbf{H}_1 = \overline{\mathbf{S}}_1^{(r)}\mathbf{V}_1^{(r)T} \; ; \; \mathbf{G}_1 = \frac{1}{\sqrt{\gamma_i}}\mathbf{U}_{1,i}\begin{bmatrix} \mathbf{I}_{r\times r} \\ 0_{|\mathcal{X}|-r\times r} \end{bmatrix} \; ; \; \mathbf{F} = \mathbf{B}$$

where $\overline{\mathbf{S}}_1^{(r)}$ *is* $r \times r$ *matrix with* r *largest singular values, and* $\mathbf{V}_1^{(r)}$ *is* $|\mathcal{X}| \times r$ *matrix with the corresponding right singular vectors. The same will be done for the other concatenated matrices for different* j *values and the optimal matrices are*

$$\mathbf{H}_j = \overline{\mathbf{S}}_j^{(r)}\mathbf{V}_j^{(r)T} \; ; \; \mathbf{G}_j = \frac{1}{\sqrt{\gamma_{i-j+1}}}\mathbf{U}_{j,i-j+1}\begin{bmatrix} \mathbf{I}_{r\times r} \\ 0_{|\mathcal{X}|-r\times r} \end{bmatrix}$$

Theorem 3.5 is a powerful theorem where it shows how to solve a multiobjective low rank approximation problem which is of crucial importance for our constrained information exchange problem.

Proof. Without loss of generality, we will use the case where $j = 1$. The eigenvalue decomposition of the concatenated matrices in **??**

$$\sqrt{\gamma_{i-j+1}}\mathbf{B}^j = \sqrt{\gamma_{i-j+1}}\mathbf{Q}_B\mathbf{\Delta}_B^j\mathbf{Q}_B^{-1} \; ; \; j \in \{1, \cdots, i\}$$

The singular value decomposition based on the concatenated matrices

$$\begin{aligned}\sqrt{\gamma_1}\mathbf{B}^i &= \mathbf{U}_{1,1}\overline{\mathbf{S}}_1\mathbf{V}_1^T\\ &\vdots\\ \sqrt{\gamma_i}\mathbf{B} &= \mathbf{U}_{1,i}\overline{\mathbf{S}}_1\mathbf{V}_1^T\end{aligned}$$

Then we will write the left singular vectors in terms of the $\mathbf{B}$, $\overline{\mathbf{S}}_1$, and $\mathbf{V}_1^T$.

$$\begin{aligned}\mathbf{U}_{1,1} &= \sqrt{\gamma_1}\mathbf{B}^i\mathbf{V}_1\overline{\mathbf{S}}_1^{-1}\\ &\vdots\\ \mathbf{U}_{1,i} &= \sqrt{\gamma_i}\mathbf{B}\mathbf{V}_1\overline{\mathbf{S}}_1^{-1}\\ \mathbf{U}_{1,j-1}(\mathbf{U}_{1,j})^{-1} &= \sqrt{\frac{\gamma_{j-1}}{\gamma_j}}\mathbf{B}\mathbf{V}_1\overline{\mathbf{S}}_1^{-1}\end{aligned}$$

Therefore, we realize that the left singular vectors are related in terms of a scaling factor and the $\mathbf{B}$ matrix. Therefore,

$$\mathbf{F} = \mathbf{B}$$

The $\mathbf{G}_1, \mathbf{H}_1$ will be equal to the low rank approximation of the concatenated matrix since it is the optimal approximation.

$$\begin{aligned}\mathbf{H}_1 &= \overline{\mathbf{S}}_1^{(r)}\mathbf{V}_1^{(r)T}\\ \mathbf{G}_1 &= \frac{1}{\sqrt{\gamma_i}}\mathbf{U}_{1,i}\begin{bmatrix}\mathbf{I}_{r\times r}\\ 0_{|\mathcal{X}|-r\times r}\end{bmatrix}\end{aligned}$$

Using the same singular value decomposition of the concatenated matrix for different j values, $\mathbf{G}_j$, and $\mathbf{H}_j$ will be derived. □

3.2.3 Computational and Communication Complexity

The performance of the multiple index-specific posterior inference algorithm is expected to be sub-optimal due to the constrained messages and the insufficient statistics. However, these constraints have the advantage of reducing communication complexity, which is particularly useful in modern applications where high-dimensional data is prevalent.

Table 3.2 provides a comparison of computational and communication complexity for the constrained and unconstrained message passing algorithms. As shown in the table, the computational complexity of the messages $O(|\mathcal{X}|^2)$.

In the preceding algorithm, the DTM is utilized within the context of multiple index-specific posterior inference. As a defining characteristic of the DTM model, the largest singular value is 1. This implies the necessity to incorporate all the previous observations (n), yielding a total computational complexity of $O(n|\mathcal{X}|^2)$.

Conversely, there exists an alternative approach through the use of CDM. In the CDM, the largest singular value is less than or equal to 1. This property substantiates the possibility of dismissing the observations from earlier stages as the singular values associated with the CDM product matrix will diminish towards an extremely small magnitude. The theorems, namely Theorem 3.4, and 3.5, cannot be applied due to the singularity of the CDM matrix. To utilize these theorems, it is necessary to perturb the matrix so that it becomes non-singular. Additional details are provided in Appendix B.

Therefore, to find the total computational complexity a set ($\mathcal{S}$) is constructed that includes the observation nodes utilized in each objective, taking into consideration that older observations might be dismissed due to the extremely small magnitude of the singular values. The cardinality of the set is ($|\mathcal{S}| = n_3$).

The communication complexity for $\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i}$ is reduced from $|\mathcal{X}|$ to r, which is particularly useful in the case of observation nodes with limited communication capabilities or in the case of high-dimensional data. These advantages come at the cost of a sub-optimal algorithm performance when compared to the case of unconstrained messages.

Despite the sub-optimality of the algorithm, it remains a practical and useful solution for high-dimensional data problems, particularly when communication resources are limited. This scenario highlights the importance of considering the trade-off between algorithm performance and resource utilization in modern data-driven applications.

		Constrained Message Passing	Unconstrained Message Passing
Computational Complexity	$\mathbf{m}_{\mathsf{x}_i\to\mathsf{x}_{i+1}}$	$O(\lvert\mathcal{X}\rvert^2)$	$O(\lvert\mathcal{X}\rvert^2)$
	Total	$O(n_2\lvert\mathcal{X}\rvert^2)$	$O(n\lvert\mathcal{X}\rvert^2)$
Communication Complexity	$\mathbf{m}_{\mathsf{y}_i\to\mathsf{x}_i}$	$O(r)$	$O(\lvert\mathcal{X}\rvert)$
	$\mathbf{m}_{\mathsf{x}_i\to\mathsf{x}_{i+1}}$	$O(\lvert\mathcal{X}\rvert)$	$O(\lvert\mathcal{X}\rvert)$

Table 3.2: Computational and communication complexity comparison between the constrained and unconstrained message passing algorithm in multiple index-specific posterior inference. n is the number of observation used for the posterior predication and $n_2 = \min\{n, n_3\}$

3.2.4 Numerical Results

This section aims to explore the performance of the constrained information exchange message passing algorithm in multiple index-specific posterior inference, using synthetic dataset. The focus will be on the case $D(\mathbf{p}_{\mathsf{x}_3|\mathsf{y}_1^2}||\hat{\mathbf{p}}_{\mathsf{x}_3|\mathsf{y}_1^2})$ and $D(\mathbf{p}_{\mathsf{x}_4|\mathsf{y}_1^3}||\hat{\mathbf{p}}_{\mathsf{x}_4|\mathsf{y}_1^3})$ as illustrated in Fig. 3-8.

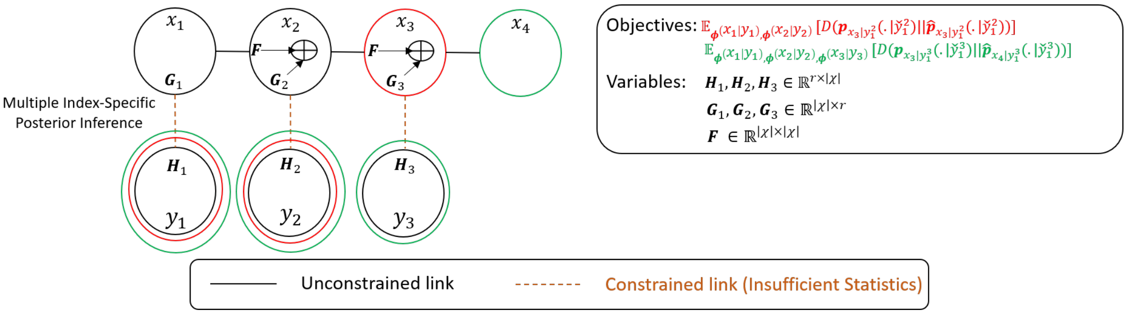

x_1
x_2
x_3
x_4
G_1
G_2
G_3
F
F
Multiple Index-Specific Posterior Inference
H_1
H_2
H_3
y_1
y_2
y_3
Objectives: $\mathbb{E}_{\phi(x_1|y_1),\phi(x_2|y_2)}[D(\boldsymbol{p}_{x_3|y_1^2}(.|\check{y}_1^2)||\hat{\boldsymbol{p}}_{x_3|y_1^2}(.|\check{y}_1^2))]$
$\mathbb{E}_{\phi(x_1|y_1),\phi(x_2|y_2),\phi(x_3|y_3)}[D(\boldsymbol{p}_{x_3|y_1^3}(.|\check{y}_1^3)||\hat{\boldsymbol{p}}_{x_4|y_1^3}(.|\check{y}_1^3))]$
Variables: $\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3 \in \mathbb{R}^{r\times|\mathcal{X}|}$
$\boldsymbol{G}_1,\boldsymbol{G}_2,\boldsymbol{G}_3 \in \mathbb{R}^{|\mathcal{X}|\times r}$
$\boldsymbol{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$
Unconstrained link
Constrained link (Insufficient Statistics)

Figure 3-8: Constrained Information Exchange Message-Passing Algorithm in Multiple Index-Specific Posterior Inference

3.2.4.1 Synthetic Dataset

Multiobjective Optimization Tradeoff: Fig. 3-9 exhibits the intrinsic tradeoff encountered in multiple index-specific posterior inference due to the challenge of transmitting sufficient statistics. The fundamental issue at hand is that the optimal solution for one objective does not invariably equate to the optimal solution for the weighted sum. The optimal solution for the combined objectives aligns with the first objective when the weighting coefficient, γ, is assigned a value of 1, thereby fully prioritizing the first objective. In contrast, a γ value of 0 ensures that the optimal solution for the weighted sum adheres to the second objective, thereby focusing exclusively on this latter objective. This interplay of weight assignment, as determined by γ, shows the inherent tradeoff between the two different objectives. Adjusting γ provides a method to explore the array of potential optimal solutions, and its value must be precisely selected, considering the particular requirements and constraints of the specific problem scenario.

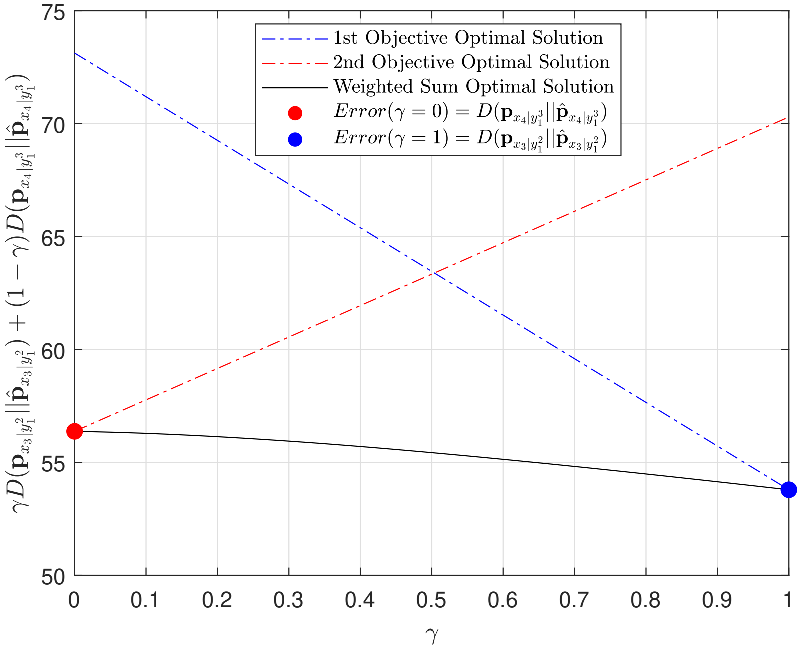


Figure 3-9: Multiobjective Optimization Tradeoff

Dependency of Error on Link Size: Fig. 3-10 illustrates the dependency between the error and the link size. As observed, there is a consistent decrease in error with an increase in the link size. This trend is attributed to the larger volume of information exchanged as the link size expands. Ultimately, the error drops to zero when the link size equals the rank of the DTM matrix.

Determining Link Sizes: Guided by a Predetermined Error Threshold: Fig. 3-11 Fig. 3-11 examines a scenario in which the objective is to limit the error below a preset threshold. This situation calls for a strategic choice of link size that aligns with this error constraint. As illustrated in Figure 3-11, maintaining the error below the 5 threshold demands a minimum link size of 154 ($r \geq 154$). Therefore, while maintaining the error threshold the reduction in the size of the link is by almost **25%**. This emphasizes the significance of constrained information message passing algorithm to deal with high-dimensional data.

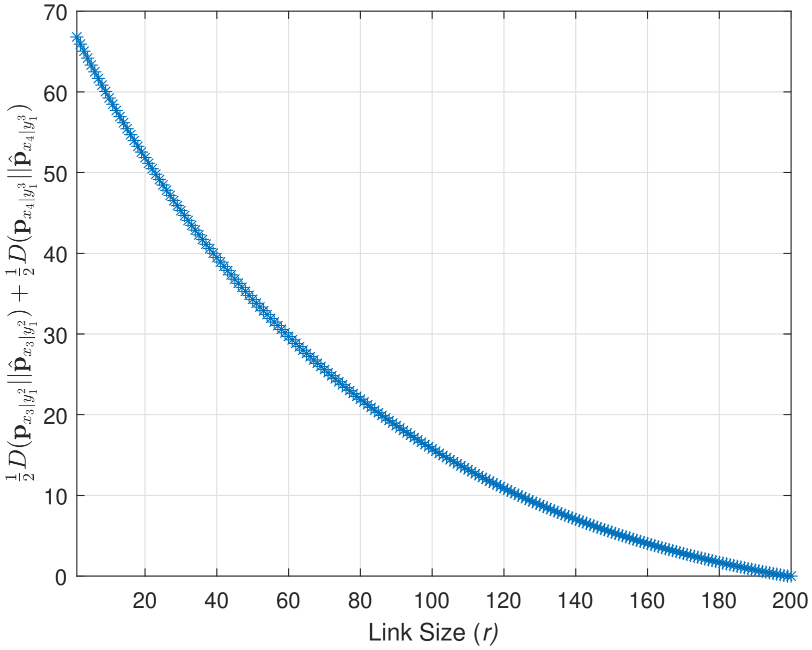


Figure 3-10: Multiple Index-Specific Posterior Inference error as a function of link size

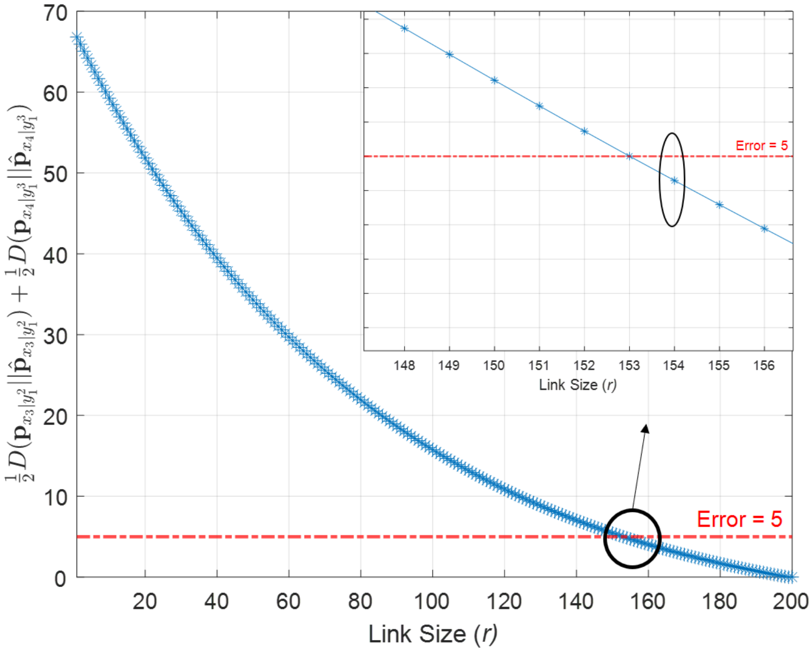


Figure 3-11: Link Sizes Selection Based on Desired Error Threshold in the Multiple Index-Specific Posterior Inference

Link Size Selection: Shaped by Error Threshold and Maximum Size Constraints: Fig. 3-12 illustrates the scenarios where the task is to maintain the error below a particular threshold, whilst staying within specific link size constraints. Such circumstances call for an intricate balance between these two defining parameters. As depicted in Fig. 3-12, if the target is to cap the error at less than 5, with the link size not surpassing 155, the suitable range for link size should fall within 154 and 155 ($154 \leq r \leq 155$). This underscores the significance of precise link size selection, which is shaped by both the desired error threshold and the maximum size limitations.

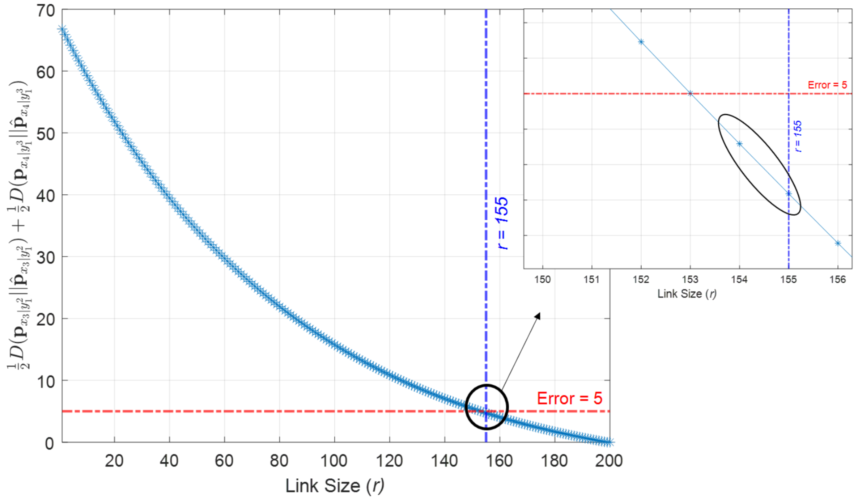


Figure 3-12: Link Size Selection Based on Desired Error Threshold and Maximum Size Constraint in the Multiple Index-Specific Posterior Inference

Error Threshold Impact on Minimum Link Size: Fig. 3-13 illustrate the trade-off between the error threshold and the corresponding minimum link size. As observed, there is a consistent decrease in the minimum link size with an increase in the associated error. This trend is attributed to the smaller amount of information exchanged as the link size decreases. Fig. 3-13 emphasizes the balance between optimizing resource utilization and achieving the necessary precision. As the size of the link decreases, the system becomes more efficient in terms of resource usage. However, it also necessitates a higher error, potentially affecting the overall performance and accuracy of the system. The figure presents a valuable perspective on system design considerations in the constrained information exchange message passing algorithm in high dimensional data.

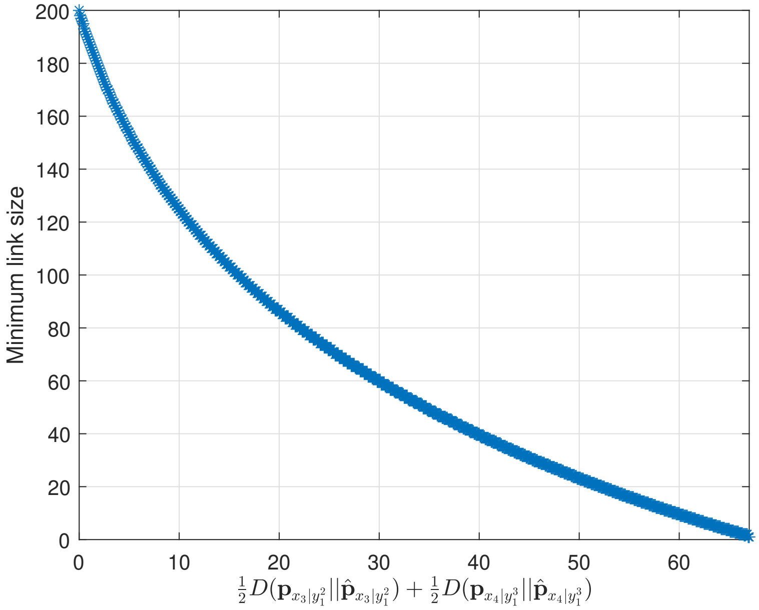


Figure 3-13: Error Threshold Impact on Minimum Link Size

Chapter 4

Index-Free Posterior Inference

This chapter delves into the domain of posterior inference, specifically focusing on index-free posteriors. In the preceding chapter, we explored index-specific posterior inference, where nodes were examined in scenarios involving varying compression matrices. Now, our attention turns to index-free posterior inference, which ensures a crucial element of uniformity by employing identical compression matrices across all nodes. This unique characteristic adds a captivating dimension to posterior inference, unveiling new insights and opening doors to diverse opportunities for analysis.

The primary objective of this chapter is to unravel the intricacies associated with index-free posterior inference, providing a comprehensive understanding of its underlying principles, practical applications, and potential benefits in the realm of statistical inference. Throughout our exploration, we will navigate the theoretical foundations and delve into the complexities of index-free posterior inference, shedding light on its various aspects.

- Single-Step Index-Free Posterior Inference: In this scenario, the primary emphasis lies in performing posterior inference based on the most recent observation. The goal is to identify the optimal matrices for compressing messages. Unlike the index-specific approach, where nodes may have different compression matrices, the index-free approach ensures that all nodes employ identical compression matrices.

- Multi-Step Index-Free Posterior Inference: In this scenario, the primary emphasis lies in performing posterior inference at multiple indexes, revealing the inherent tradeoff nature of the problem. Optimization plays a crucial role in achieving a delicate balance between various posterior inference objectives. By solving the optimization problem, optimal compression matrices are obtained and effectively utilized for message compression. Unlike the multiple index-specific posterior inference approach, the multi-step index-free posterior inference employs identical compression matrices across all nodes, making it more feasible and applicable in real-world scenarios.

4.1 Single Step Index-Free Posterior Inference

This scenario involves estimation the posterior distribution using the last observation. In the index-free approach, the compression matrices will be the same among all nodes, which is not the case in index-specific, as depicted in Fig. 4-1.

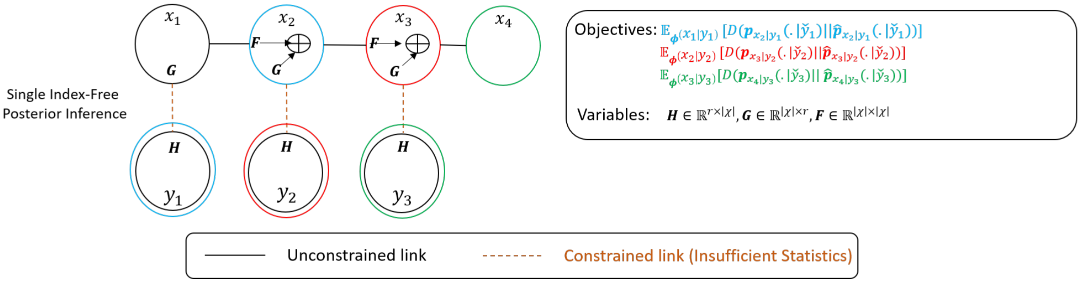

x_1
x_2
x_3
x_4
F
G
H
y_1
y_2
y_3
Single Index-Free Posterior Inference
Objectives: $\mathbb{E}_{\phi(x_1|y_1)}[D(\boldsymbol{p}_{x_2|y_1}(.|\check{y}_1)||\hat{\boldsymbol{p}}_{x_2|y_1}(.|\check{y}_1))]$
$\mathbb{E}_{\phi(x_2|y_2)}[D(\boldsymbol{p}_{x_3|y_2}(.|\check{y}_2)||\hat{\boldsymbol{p}}_{x_3|y_2}(.|\check{y}_2))]$
$\mathbb{E}_{\phi(x_3|y_3)}[D(\boldsymbol{p}_{x_4|y_3}(.|\check{y}_3)||\hat{\boldsymbol{p}}_{x_4|y_3}(.|\check{y}_3))]$
Variables: $\boldsymbol{H} \in \mathbb{R}^{r\times|\mathcal{X}|}, \boldsymbol{G} \in \mathbb{R}^{|\mathcal{X}|\times r}, \boldsymbol{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$
Unconstrained link
Constrained link (Insufficient Statistics)

Figure 4-1: Constrained Information Exchange Message-Passing Algorithm in Single-Step Index-Free Posterior Inference

4.1.1 Fixed Divergence Transfer Matrix

In this section $\mathbf{B} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ is known exactly and is fixed. The messages are constrained $\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i} \in \mathbb{R}^r$, where $r < |\mathcal{X}|$.

$$\mathbf{B}_{x_i,x_{i+1}} = \mathbf{B} \ ; \ \forall i$$

- at node y_i, pick a matrix $\mathbf{H} \in \mathbb{R}^{r\times|\mathcal{X}|}$, to reduce the size of the message ($\mathbf{m}_{\mathsf{y}_i\to\mathsf{x}_i}$) to r-dimensional message ($\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i}$) since there is a constrained link between y_i and x_i.
$$\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i} = \mathbf{H}\mathbf{m}_{\mathsf{y}_i\to\mathsf{x}_i} = \mathbf{H}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$$

- at node x_i, pick a matrix $\mathbf{G} \in \mathbb{R}^{|\mathcal{X}|\times r}$ and $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ to calculate the forward message
$$\hat{\mathbf{m}}_{\mathsf{x}_i\to\mathsf{x}_{i+1}} = \mathbf{G}\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i} = \mathbf{G}\mathbf{H}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)} = \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)} = \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)}$$

- Optimization:
$$\begin{aligned}
&\min_{\mathbf{H},\mathbf{G}} \;\; \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}}\left[D(\mathbf{p}_{\mathsf{x}_{i+1}|\mathsf{y}_i}(.|\check{\mathsf{y}}_i)||\hat{\mathbf{p}}_{\mathsf{x}_{i+1}|\mathsf{y}_i}(.|\check{\mathsf{y}}_i))\right]\\
&= \min_{\mathbf{H},\mathbf{G}} \;\; \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}}\left[\frac{\epsilon^2}{2}\left\|\boldsymbol{\phi}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)} - \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)}\right\|_2^2\right] \text{ (Using lemma 2.7)}\\
&= \min_{\mathbf{H},\mathbf{G}} \;\; \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}}\left[\frac{\epsilon^2}{2}\left\|\mathbf{B}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)} - \mathbf{G}\mathbf{H}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}\right\|_F^2\right]
\end{aligned}$$

Using lemma 3.1 and assuming $\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$ are spherical symmetric
$$\begin{aligned}
&= \min_{\mathbf{H},\mathbf{G}} \;\; \left[\frac{\epsilon^2}{2|\mathcal{X}|}\,\|(\mathbf{B}-\mathbf{G}\mathbf{H})\|_F^2\,\mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}}\left[\left\|\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}\right\|_2^2\right]\right]\\
&= \frac{\alpha\epsilon^2}{2|\mathcal{X}|}\min_{\mathbf{H},\mathbf{G}} \;\; \left[\|(\mathbf{B}-\mathbf{G}\mathbf{H})\|_F^2\right] \;\; ; \;\mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}}\left[\left\|\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}\right\|_2^2\right] = \alpha
\end{aligned}$$

By using theorem 3.2, the optimal $\mathbf{G}, \mathbf{H}$ matrices
$$\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{V}^T \implies \mathbf{G} = \mathbf{U}^{(r)} \; ; \; \mathbf{H} = \mathbf{S}_j^{(r)}(\mathbf{V}_j^{(r)})^T.$$

Even if we optimize over different i, the optimal matrices will be the same and this is because all of the objectives are the same and have the same DTM matrix.

- at node x_{i+1}, we will calculate the posterior distribution

$$\hat{p}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(x_{i+1}|\check{y}_1^i) = p_{\mathsf{x}_{i+1}}(x_{i+1}) + \epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}\hat{m}_{\mathsf{x}_i\to\mathsf{x}_{i+1}}(x_{i+1})$$

4.1.2 Computation and Communication Complexity

The performance of the single step index-free posterior inference is expected to be sub-optimal due to the constrained messages and the insufficient statistics. However, these constraints have the advantage of reducing computational and communication complexity, which is particularly useful in modern applications where high-dimensional data is prevalent.

Table 4.1 provides a comparison of computational and communication complexity for the constrained and unconstrained message passing algorithms. As shown in the table, the computational complexity is reduced from $O(|\mathcal{X}|^2)$ to $O(r|\mathcal{X}|)$, where $r < |\mathcal{X}|$. This reduction in complexity is significant, particularly for models with large states, such as language or vision models.

Furthermore, the communication complexity for $\mathbf{m}_{\mathsf{y}_i\to\mathsf{x}_i}$is reduced from $O(|\mathcal{X}|)$ to $O(r)$, which is particularly useful in the case of observation nodes with limited communication capabilities. These advantages come at the cost of a sub-optimal algorithm performance when compared to the case of unconstrained messages.

Despite the sub-optimality of the algorithm, it remains a practical and useful solution for high-dimensional data problems, particularly when computational and communication resources are limited. This scenario highlights the importance of considering the trade-off between algorithm performance and resource utilization in modern data-driven applications.

		Constrained Message Passing	Unconstrained Message Passing
Computational Complexity	$\mathbf{m}_{\mathsf{x}_i \to \mathsf{x}_{i+1}}$	$O(r\|\mathcal{X}\|)$	$O(\|\mathcal{X}\|^2)$
	Total	$O(nr\|\mathcal{X}\|)$	$O(n\|\mathcal{X}\|^2)$
Communication Complexity	$\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i}$	$O(r)$	$O(\|\mathcal{X}\|)$
	$\mathbf{m}_{\mathsf{x}_i \to \mathsf{x}_{i+1}}$	$O(\|\mathcal{X}\|)$	$O(\|\mathcal{X}\|)$

Table 4.1: Computational and communication complexity comparison between the constrained and unconstrained message passing algorithm in single step index-free posterior inference. n is the number of observation used for the posterior predication.

4.1.3 Numerical Results

This section aims to explore the performance of the constrained information exchange message passing algorithm in single-step index-free posterior inference, using synthetic dataset. The focus will be on the case $\sum_{j=1}^{10} D(\mathbf{p}_{\mathsf{x}_j|\mathsf{y}_{j-1}}||\hat{\mathbf{p}}_{\mathsf{x}_j|\mathsf{y}_{j-1}})$ as illustrated in Fig. 4-8.

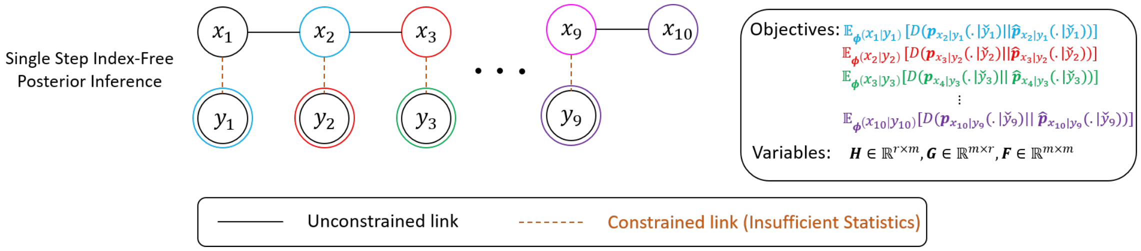

Single Step Index-Free Posterior Inference
x_1
x_2
x_3
x_9
x_{10}
y_1
y_2
y_3
y_9
Objectives: $\mathbb{E}_{\phi(x_1|y_1)}[D(\boldsymbol{p}_{x_2|y_1}(.|\check{y}_1)||\hat{\boldsymbol{p}}_{x_2|y_1}(.|\check{y}_1))]$
$\mathbb{E}_{\phi(x_2|y_2)}[D(\boldsymbol{p}_{x_3|y_2}(.|\check{y}_2)||\hat{\boldsymbol{p}}_{x_3|y_2}(.|\check{y}_2))]$
$\mathbb{E}_{\phi(x_3|y_3)}[D(\boldsymbol{p}_{x_4|y_3}(.|\check{y}_3)||\hat{\boldsymbol{p}}_{x_4|y_3}(.|\check{y}_3))]$
$\vdots$
$\mathbb{E}_{\phi(x_{10}|y_{10})}[D(\boldsymbol{p}_{x_{10}|y_9}(.|\check{y}_9)||\hat{\boldsymbol{p}}_{x_{10}|y_9}(.|\check{y}_9))]$
Variables: $\boldsymbol{H} \in \mathbb{R}^{r\times m}, \boldsymbol{G} \in \mathbb{R}^{m\times r}, \boldsymbol{F} \in \mathbb{R}^{m\times m}$
Unconstrained link
Constrained link (Insufficient Statistics)

Figure 4-2: Constrained Information Exchange Message-Passing Algorithm in Single-Step Index-Free Posterior Inference

4.1.3.1 Synthetic Dataset

Dependency of Error on Link Size: Fig. 4-3 illustrates the dependency between the error and the link size. As observed, there is a consistent decrease in error with an increase in the link size. This trend is attributed to the larger volume of information exchanged as the link size expands. Ultimately, the error drops to zero when the link size equals the rank of the DTM matrix.

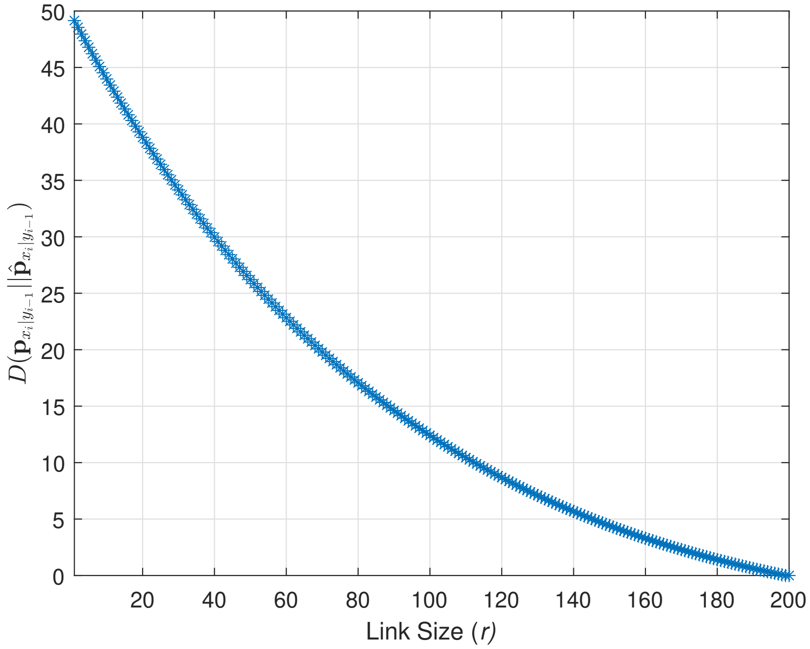


Figure 4-3: Single Step Index-Free Posterior Inference error as a function of link size

Determining Link Sizes: Guided by a Predetermined Error Threshold: Fig. 4-4 examines a scenario in which the objective is to limit the error below a preset threshold. This situation calls for a strategic choice of link size that aligns with this error constraint. As illustrated in Figure 4-4, maintaining the error below the 5 threshold demands a minimum link size of 146 ($r \geq 146$). Therefore, while maintaining the error threshold the reduction in the size of the link is by almost **27%**. This emphasizes the significance of constrained information message passing algorithm to deal with high-dimensional data.

Link Size Selection: Shaped by Error Threshold and Maximum Size Constraints: Fig. 4-5 illustrates the scenarios where the task is to maintain the error below a threshold, while staying within specific link size constraints. Such circumstances require balance between these two objectives. As depicted in Fig. 4-5, if the target is to cap the error at less than 5, with the link size not surpassing 155, the suitable range for link size ($146 \leq r \leq 155$). This shows the significance of link size selection, which is shaped by the error threshold and the maximum size limitations.

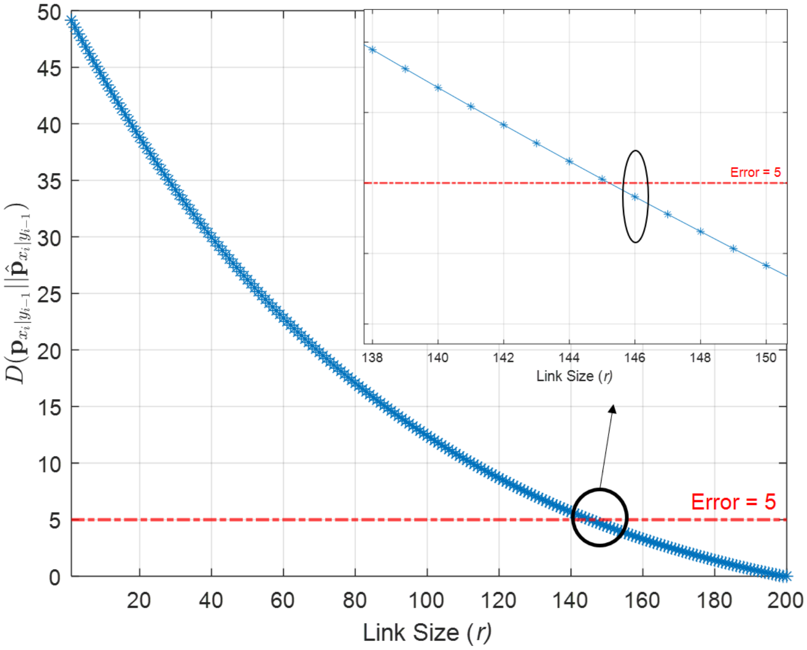


Figure 4-4: Link Sizes Selection Based on Desired Error Threshold in the Single-Step Index-Free Posterior Inference

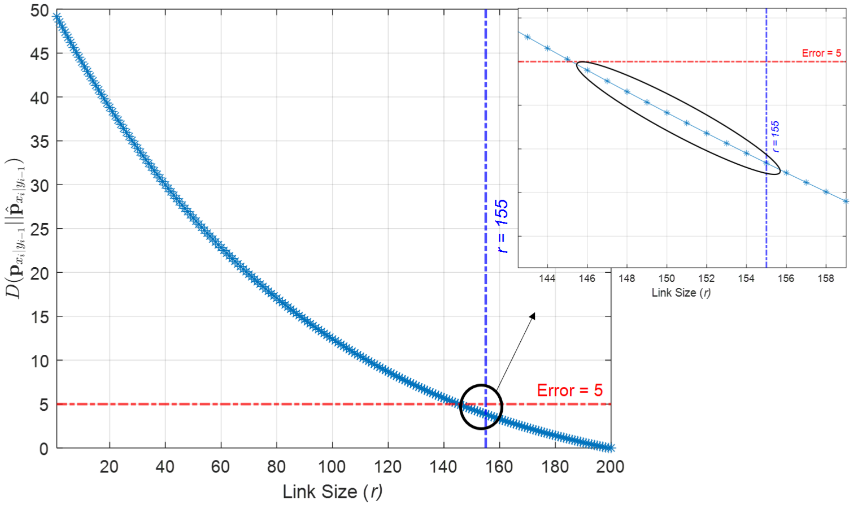


Figure 4-5: Link Size Selection Based on Desired Error Threshold and Maximum Size Constraint in the Single-Step Index-Free Posterior Inference

Error Threshold Impact on Minimum Link Size: Fig. 4-6 illustrate the trade-off between the error threshold and the corresponding minimum link size. As observed, there is a consistent decrease in the minimum link size with an increase in the associated error. This trend is attributed to the smaller amount of information exchanged as the link size decreases. Fig. 4-6 emphasizes the balance between optimizing resource utilization and achieving the necessary precision. As the size of the link decreases, the system becomes more efficient in terms of resource usage. However, it also necessitates a higher error, potentially affecting the overall performance and accuracy of the system. The figure presents a valuable perspective on system design considerations in the constrained information exchange message passing algorithm in high dimensional data.

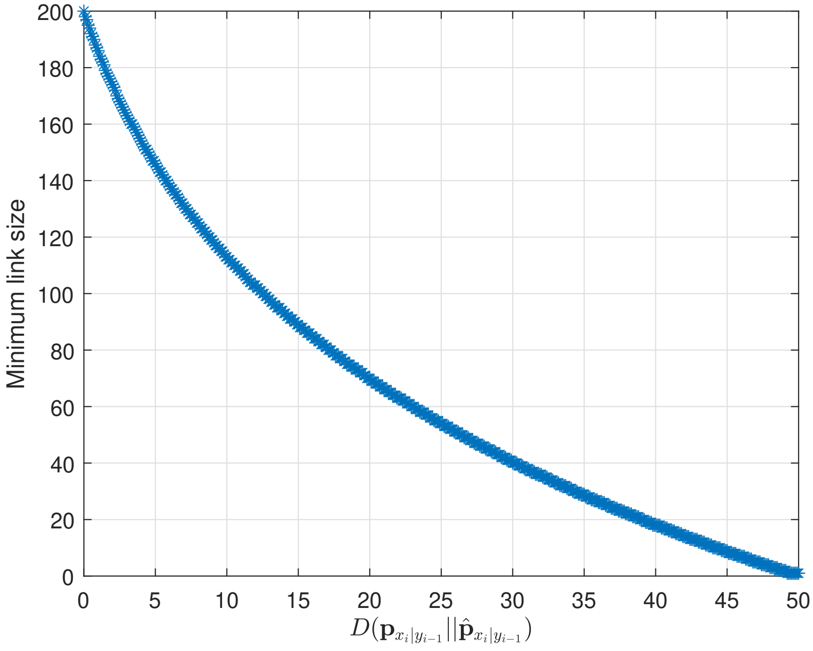


Figure 4-6: Error Threshold Impact on Minimum Link Size

4.2 Multi-Step Index-Free Posterior Inference

In this particular scenario, the task at hand is to estimate the posterior distribution by utilizing past observations. Notably, all nodes within the system employ identical compression matrices, which differentiates it from the index-specific approach. The process involves solving a multi-objective optimization problem to identify the optimal matrices for compressing messages, as depicted in Fig. 4-7. This scenario sheds light on the inherent tradeoff nature of the problem, which stems from the challenge of transmitting insufficient statistics.

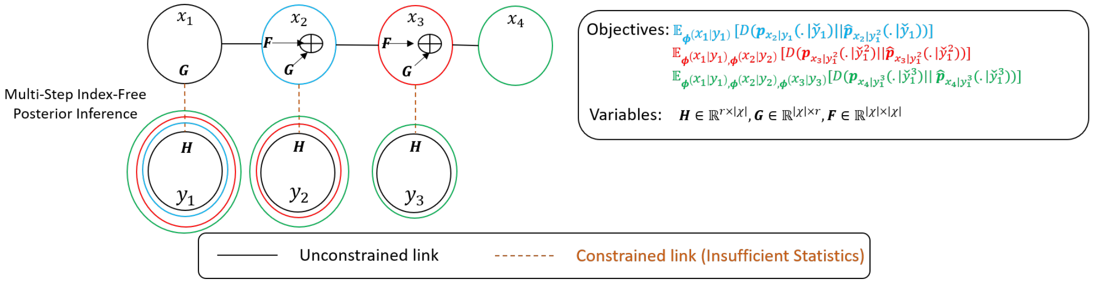

x_1
x_2
x_3
x_4
F
G
H
y_1
y_2
y_3
Multi-Step Index-Free Posterior Inference
Objectives: $\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1)}[D(\boldsymbol{p}_{x_2|y_1}(.|\check{y}_1)||\hat{\boldsymbol{p}}_{x_2|y_1^2}(.|\check{y}_1))]$
$\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2)}[D(\boldsymbol{p}_{x_3|y_1^2}(.|\check{y}_1^2)||\hat{\boldsymbol{p}}_{x_3|y_1^2}(.|\check{y}_1^2))]$
$\mathbb{E}_{\boldsymbol{\phi}(x_1|y_1),\boldsymbol{\phi}(x_2|y_2),\boldsymbol{\phi}(x_3|y_3)}[D(\boldsymbol{p}_{x_4|y_1^3}(.|\check{y}_1^3)||\hat{\boldsymbol{p}}_{x_4|y_1^3}(.|\check{y}_1^3))]$
Variables: $\boldsymbol{H} \in \mathbb{R}^{r\times|\chi|}, \boldsymbol{G} \in \mathbb{R}^{|\chi|\times r}, \boldsymbol{F} \in \mathbb{R}^{|\chi|\times|\chi|}$
Unconstrained link
Constrained link (Insufficient Statistics)

Figure 4-7: Constrained Information Exchange Message-Passing Algorithm in Multi-Step Index-Free Posterior Inference

4.2.1 Fixed Divergence Transfer Matrix

In this section $\mathbf{B} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ is known exactly and is fixed. The messages are constrained $\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i} \in \mathbb{R}^r$, where $r < |\mathcal{X}|$.

- at node y_i, pick a matrix $\mathbf{H} \in \mathbb{R}^{r\times|\mathcal{X}|}$, to reduce the size of the message $(\mathbf{m}_{\mathsf{y}_i\to\mathsf{x}_i})$ to r-dimensional message $(\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i})$ since there is a constrained link between y_i and x_i.

$$\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i} = \mathbf{H}\mathbf{m}_{\mathsf{y}_i\to\mathsf{x}_i} = \mathbf{H}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}$$

- at node x_i, pick a matrix $\mathbf{G} \in \mathbb{R}^{|\mathcal{X}|\times r}$ and $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$ to calculate the forward message

$$\hat{\mathbf{m}}_{\mathsf{x}_i\to\mathsf{x}_{i+1}} = \mathbf{G}\hat{\mathbf{m}}_{\mathsf{y}_i\to\mathsf{x}_i} + \mathbf{F}\hat{\mathbf{m}}_{\mathsf{x}_{i-1}\to\mathsf{x}_i}$$

$$
\begin{aligned}
&= \mathbf{GH}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)} + \mathbf{F}\hat{\boldsymbol{\phi}}^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})} \\
&= \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_i)} + \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^{i-1})} \\
&= \hat{\boldsymbol{\phi}}^{(\mathsf{x}_{i+1}|\mathsf{y}_1^{i})}
\end{aligned}
$$

- Optimization objectives

$$
\begin{aligned}
&\min_{\mathbf{H},\mathbf{G},\mathbf{F}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}} \left[D(\mathbf{p}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(.|\check{\mathsf{y}}_1^i)||\hat{\mathbf{p}}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(.|\check{\mathsf{y}}_1^i))\right] \\
&\min_{\mathbf{H},\mathbf{G},\mathbf{F}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)},\cdots,\boldsymbol{\phi}^{(\mathsf{x}_{i-1}|\mathsf{y}_{i-1})}} \left[D(\mathbf{p}_{\mathsf{x}_{i}|\mathsf{y}_1^{i-1}}(.|\check{\mathsf{y}}_1^{i-1})||\hat{\mathbf{p}}_{\mathsf{x}_{i}|\mathsf{y}_1^{i-1}}(.|\check{\mathsf{y}}_1^{i-1}))\right] \\
&\qquad\qquad\qquad\qquad \vdots \\
&\min_{\mathbf{H},\mathbf{G}} \quad \mathbb{E}_{\boldsymbol{\phi}^{(\mathsf{x}_1|\mathsf{y}_1)}} \left[D(\mathbf{p}_{\mathsf{x}_2|\mathsf{y}_1}(.|\check{\mathsf{y}}_1)||\hat{\mathbf{p}}_{\mathsf{x}_2|\mathsf{y}_1}(.|\check{\mathsf{y}}_1))\right]
\end{aligned}
$$

- Using lemma 2.7 and 3.1, and assuming that the information vectors are spherical symmetric and independent

$$
\begin{aligned}
&\min_{\mathbf{H},\mathbf{G},\mathbf{F}} \quad \frac{\alpha\epsilon^2}{2m}\sum_{j=1}^{i}\left\|\mathbf{B}^{i-j+1} - \mathbf{F}^{i-j}\mathbf{GH}\right\|_F^2 \\
&\min_{\mathbf{H},\mathbf{G},\mathbf{F}} \quad \frac{\alpha\epsilon^2}{2m}\sum_{j=1}^{i-1}\left\|\mathbf{B}^{i-j} - \mathbf{F}^{i-j-1}\mathbf{GH}\right\|_F^2 \\
&\qquad\qquad\qquad\qquad \vdots \\
&\min_{\mathbf{H},\mathbf{G}} \quad \frac{\alpha\epsilon^2}{2m}\left\|\mathbf{B} - \mathbf{GH}\right\|_F^2
\end{aligned}
$$

- We will write these objective functions as a weighted sum and then we will solve the weighted sum

$$
\begin{aligned}
\frac{\alpha\epsilon^2}{2m}\min_{\mathbf{H},\mathbf{G},\mathbf{F}} \quad &\Bigg[\gamma_1\sum_{j=1}^{i}\left\|\mathbf{B}^{i-j+1} - \mathbf{F}^{i-j}\mathbf{GH}\right\|_F^2 + \gamma_2\sum_{j=1}^{i-1}\left\|\mathbf{B}^{i-j} - \mathbf{F}^{i-j-1}\mathbf{GH}\right\|_F^2 \\
&+ \cdots + \gamma_i\left\|\mathbf{B} - \mathbf{GH}\right\|_F^2\Bigg] \\
\text{s.t} \qquad &\sum_i \gamma_i = 1
\end{aligned}
$$

$$\frac{\alpha\epsilon^2}{2m}\min_{\mathbf{H},\mathbf{G},\mathbf{F}} \left\| \begin{bmatrix} \sqrt{\gamma_1}\left(\mathbf{B}^{i-j+1}-\mathbf{F}^{i-j}\mathbf{G}\mathbf{H}\right) \\ \sqrt{\gamma_1+\gamma_2}\left(\mathbf{B}^{i-j}-\mathbf{F}^{i-j-1}\mathbf{G}\mathbf{H}\right) \\ \vdots \\ \sqrt{\sum_i \gamma_i}\left(\mathbf{B}-\mathbf{G}\mathbf{H}\right) \end{bmatrix} \right\|_F^2$$

- To solve this multiobjective optimization we will use the following theorem:

  Theorem 4.1. *Let* $\mathbf{B} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$, $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$,$\mathbf{G} \in \mathbb{R}^{|\mathcal{X}|\times r}$, *and* $\mathbf{H} \in \mathbb{R}^{r\times|\mathcal{X}|}$. *The SVD of the concatenated matrix*

$$\left\| \begin{bmatrix} \sqrt{\gamma_1}\mathbf{B}^i \\ \sqrt{\gamma_1+\gamma_2}\mathbf{B}^{i-1} \\ \vdots \\ \sqrt{\sum_i \gamma_i}\mathbf{B} \end{bmatrix} \right\|_F^2 = \begin{bmatrix} \underbrace{\mathbf{U}_{1,1}}_{|\mathcal{X}|\times|\mathcal{X}|} & \\ \underbrace{\mathbf{U}_{1,2}}_{|\mathcal{X}|\times|\mathcal{X}|} & \underbrace{\mathbf{G}}_{(i)|\mathcal{X}|\times|\mathcal{X}|} \\ \vdots & \\ \underbrace{\mathbf{U}_{1,i}}_{|\mathcal{X}|\times|\mathcal{X}|} & \end{bmatrix} \begin{bmatrix} \underbrace{\overline{\mathbf{S}}_1}_{|\mathcal{X}|\times|\mathcal{X}|} \\ \mathbf{0}_{(i-1)|\mathcal{X}|\times|\mathcal{X}|} \end{bmatrix} \underbrace{\mathbf{V}_1^T}_{|\mathcal{X}|\times|\mathcal{X}|}$$

  where $\mathbf{U}_{1,t}$ *and* $\mathbf{V}_1$ *are the left and right singular vectors, and* $\overline{\mathbf{S}}_1$ *are the singular values.*

  The optimal matrices of

$$\left\| \begin{bmatrix} \sqrt{\gamma_1}\left(\mathbf{B}^{i}-\mathbf{F}^{i-1}\mathbf{G}\mathbf{H}\right) \\ \sqrt{\gamma_1+\gamma_2}\left(\mathbf{B}^{i-1}-\mathbf{F}^{i-2}\mathbf{G}\mathbf{H}\right) \\ \vdots \\ \sqrt{\sum_i \gamma_i}\left(\mathbf{B}-\mathbf{G}\mathbf{H}\right) \end{bmatrix} \right\|_F^2$$

  are

$$\mathbf{H} = \overline{\mathbf{S}}_1^{(r)}\mathbf{V}_1^{(r)T} \ ; \ \mathbf{G} = \mathbf{U}_{1,i} \begin{bmatrix} \mathbf{I}_{r\times r} \\ 0_{|\mathcal{X}|-r\times r} \end{bmatrix} \ ; \ \mathbf{F} = \mathbf{B}$$

  where $\overline{\mathbf{S}}_1^{(r)}$ *is* $r \times r$ *matrix with* r *largest singular values, and* $\mathbf{V}_1^{(r)}$ *is* $|\mathcal{X}| \times r$ *matrix with the corresponding right singular vectors.*

Proof. The eigenvalue decomposition of the concatenated matrices in B.1

$$\begin{aligned}\sqrt{\gamma_1}\mathbf{B}^{i-j+1} &= \sqrt{\gamma_1}\mathbf{Q}_B\mathbf{\Delta}_B^{i-j+1}\mathbf{Q}_B^{-1}\\ &\vdots\\ \mathbf{B} &= \mathbf{Q}_B\mathbf{\Delta}_B\mathbf{Q}_B^{-1}\end{aligned}$$

The singular value decomposition based on the concatenated matrices

$$\begin{aligned}\sqrt{\gamma_1}\mathbf{B}^{i} &= \mathbf{U}_{1,1}\overline{\mathbf{S}}_1\mathbf{V}_1^T\\ &\vdots\\ \mathbf{B} &= \mathbf{U}_{1,i}\overline{\mathbf{S}}_1\mathbf{V}_1^T\end{aligned}$$

Then we will write the left singular vectors in terms of the $\mathbf{B}$, $\overline{\mathbf{S}}_1$, and $\mathbf{V}_1^T$.

$$\begin{aligned}\mathbf{U}_{1,1} &= \sqrt{\gamma_1}\mathbf{B}^i\mathbf{V}_1\overline{\mathbf{S}}_1^{-1}\\ &\vdots\\ \mathbf{U}_{1,i} &= \mathbf{B}\mathbf{V}_1\overline{\mathbf{S}}_1^{-1}\\ \mathbf{U}_{1,j-1}(\mathbf{U}_{1,j})^{-1} &= \sqrt{\frac{\sum_{t=1}^{j-1}\gamma_t}{\sum_{t=1}^{j}\gamma_t}}\mathbf{B}\mathbf{V}_1\overline{\mathbf{S}}_1^{-1} = \sqrt{\sum_{t=1}^{j-1}}\gamma_t\mathbf{B}\mathbf{V}_1\overline{\mathbf{S}}_1^{-1}\end{aligned}$$

Therefore, we realize that the left singular vectors are related in terms of a scaling factor and the $\mathbf{B}$ matrix. Therefore,

$$\mathbf{F} = \mathbf{B}$$

The $\mathbf{G}, \mathbf{H}$ will be equal to the low rank approximation of the concatenated matrix since it is the optimal approximation.

$$\mathbf{H} = \overline{\mathbf{S}}_1^{(r)}\mathbf{V}_1^{(r)T}$$

$$\mathbf{G} = \mathbf{U}_{1,i} \begin{bmatrix} \mathbf{I}_{r\times r} \\ 0_{|\mathcal{X}|-r\times r} \end{bmatrix}$$

□

- at node x_{i+1}, we will calculate the posterior distribution

$$\hat{p}_{\mathsf{x}_{i+1}|\mathsf{y}_1^i}(x_{i+1}|\check{y}_1^i) = p_{\mathsf{x}_{i+1}}(x_{i+1}) + \epsilon\sqrt{p_{\mathsf{x}_{i+1}}(x_{i+1})}\hat{m}_{\mathsf{x}_i\to\mathsf{x}_{i+1}}(x_{i+1})$$

4.2.2 Computational and Communication Complexity

The performance of the multi-step index-free posterior inference algorithm is expected to be sub-optimal due to the constrained messages and the insufficient statistics. However, these constraints have the advantage of reducing communication complexity, which is particularly useful in modern applications where high-dimensional data is prevalent.

Table 4.2 provides a comparison of computational and communication complexity for the constrained and unconstrained message passing algorithms. The communication complexity for $\mathbf{m}_{\mathsf{y}_i\to\mathsf{x}_i}$ is reduced from $|\mathcal{X}|$ to r, which is particularly useful in the case of observation nodes with limited communication capabilities or in the case of high-dimensional data. These advantages come at the cost of a sub-optimal algorithm performance when compared to the case of unconstrained messages.

As shown in the table, the computational complexity is almost the same and this is because there are no constraints on the messages $\mathbf{m}_{\mathsf{x}_i\to\mathsf{x}_{i+1}}$.

Despite the sub-optimality of the algorithm, it remains a practical and useful solution for high-dimensional data problems, particularly when communication resources are limited. This scenario highlights the importance of considering the trade-off between algorithm performance and resource utilization in modern data-driven applications.

		Constrained Message Passing	Unconstrained Message Passing
Computational Complexity	$\mathbf{m}_{\mathsf{x}_i \to \mathsf{x}_{i+1}}$	$O(\lvert\mathcal{X}\rvert^2)$	$O(\lvert\mathcal{X}\rvert^2)$
	Total	$O(n\lvert\mathcal{X}\rvert^2)$	$O(n\lvert\mathcal{X}\rvert^2)$
Communication Complexity	$\mathbf{m}_{\mathsf{y}_i \to \mathsf{x}_i}$	$O(r)$	$O(\lvert\mathcal{X}\rvert)$
	$\mathbf{m}_{\mathsf{x}_i \to \mathsf{x}_{i+1}}$	$O(\lvert\mathcal{X}\rvert)$	$O(\lvert\mathcal{X}\rvert)$

Table 4.2: Computational and communication complexity comparison between the constrained and unconstrained message passing algorithm in multi-step index-free posterior inference. n is the number of observation used for the posterior predication.

4.2.3 Numerical Results

This section aims to explore the performance of the constrained information exchange message passing algorithm in multi-step index-free posterior inference, using synthetic dataset. The focus will be on the case $\sum_{j=1}^{i} D(\mathbf{p}_{\mathsf{x}_{j+1}|\mathsf{y}_1^j} || \hat{\mathbf{p}}_{\mathsf{x}_{j+1}|\mathsf{y}_1^j})$ as illustrated in Fig. 4-8.

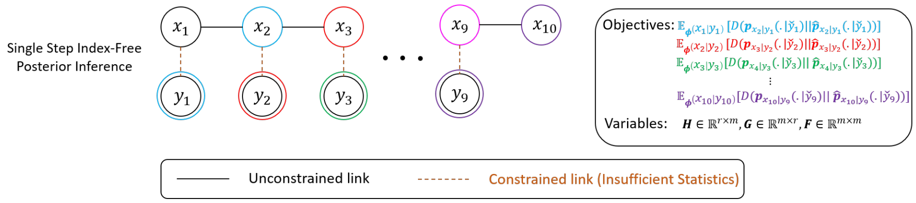

Single Step Index-Free Posterior Inference
x_1
x_2
x_3
x_9
x_{10}
y_1
y_2
y_3
y_9
. . .
Objectives: $\mathbb{E}_{\phi(x_1|y_1)}[D(\boldsymbol{p}_{x_2|y_1}(.|\check{y}_1)||\hat{\boldsymbol{p}}_{x_2|y_1}(.|\check{y}_1))]$
$\mathbb{E}_{\phi(x_2|y_2)}[D(\boldsymbol{p}_{x_3|y_2}(.|\check{y}_2)||\hat{\boldsymbol{p}}_{x_3|y_2}(.|\check{y}_2))]$
$\mathbb{E}_{\phi(x_3|y_3)}[D(\boldsymbol{p}_{x_4|y_3}(.|\check{y}_3)||\hat{\boldsymbol{p}}_{x_4|y_3}(.|\check{y}_3))]$
$\vdots$
$\mathbb{E}_{\phi(x_{10}|y_{10})}[D(\boldsymbol{p}_{x_{10}|y_9}(.|\check{y}_9)||\hat{\boldsymbol{p}}_{x_{10}|y_9}(.|\check{y}_9))]$
Variables: $\boldsymbol{H} \in \mathbb{R}^{r\times m}, \boldsymbol{G} \in \mathbb{R}^{m\times r}, \boldsymbol{F} \in \mathbb{R}^{m\times m}$
Unconstrained link
Constrained link (Insufficient Statistics)

Figure 4-8: Constrained Information Exchange Message-Passing Algorithm in Single-Step Index-Free Posterior Inference

4.2.3.1 Synthetic Dataset

Dependency of Error on Link Size: Fig. 4-9 illustrates the dependency between the error and the link size. As observed, there is a consistent decrease in error with an increase in the link size. This trend is attributed to the larger volume of information exchanged as the link size expands. Ultimately, the error drops to zero when the link size equals the rank of the DTM matrix.

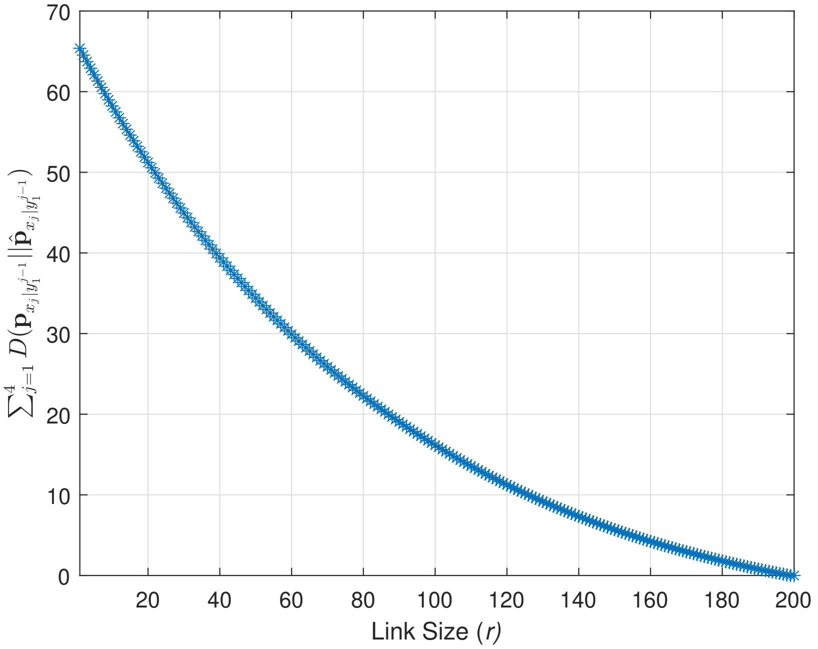


Figure 4-9: Multi-Step Index-Free Posterior Inference error as a function of link size

Determining Link Sizes: Guided by a Predetermined Error Threshold: Fig. 4-10 examines a scenario in which the objective is to limit the error below a preset threshold. This situation calls for a strategic choice of link size that aligns with this error constraint. As illustrated in Figure 4-10, maintaining the error below the 5 threshold demands a minimum link size of 156 ($r \geq 156$). Therefore, while maintaining the error threshold the reduction in the size of the link is by almost **22%**. This emphasizes the significance of constrained information message passing algorithm to deal with high-dimensional data.

Link Size Selection: Shaped by Error Threshold and Maximum Size Constraints: Fig. 4-11 illustrates the scenarios where the task is to maintain the error below a threshold, while staying within specific link size constraints. Such circumstances require balance between these two objectives. As depicted in Fig. 4-11, if the target is to cap the error at less than 5, with the link size not surpassing 160, the suitable range for link size ($156 \leq r \leq 160$). This shows the significance of link size selection, which is shaped by the error threshold and the maximum size limitations.

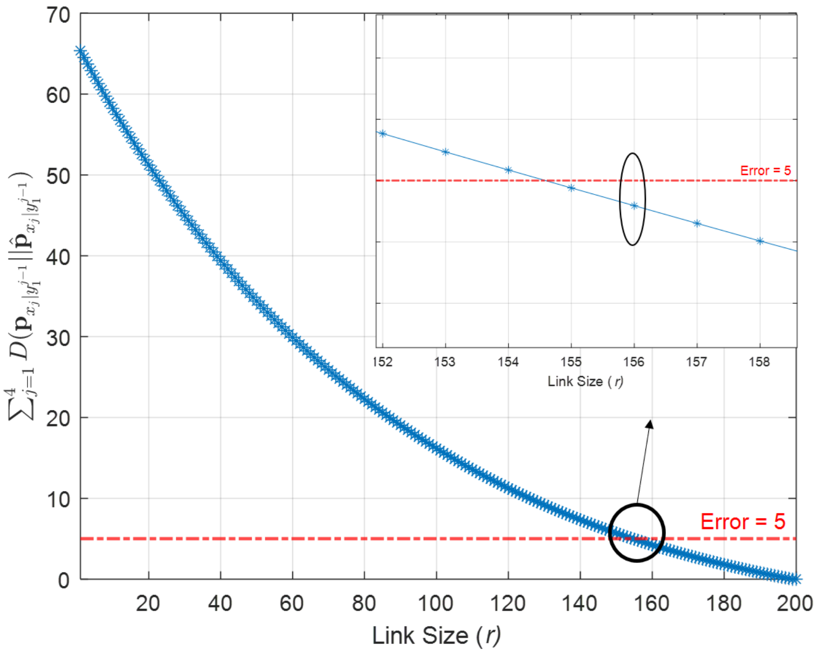


Figure 4-10: Link Sizes Selection Based on Desired Error Threshold in the Multi-Step Index-Free Posterior Inference

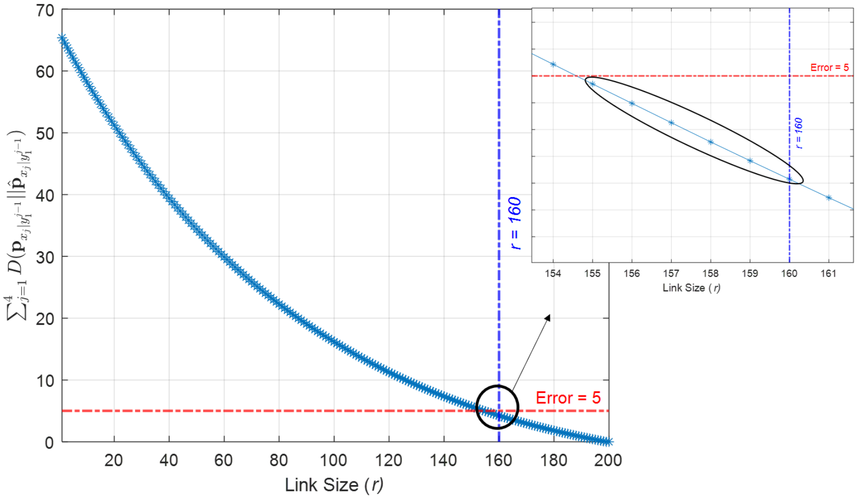


Figure 4-11: Link Size Selection Based on Desired Error Threshold and Maximum Size Constraint in the Multi-Step Index-Free Posterior Inference

Error Threshold Impact on Minimum Link Size: Fig. 4-12 illustrate the trade-off between the error threshold and the corresponding minimum link size. As observed, there is a consistent decrease in the minimum link size with an increase in the associated error. This trend is attributed to the smaller amount of information exchanged as the link size decreases. Fig. 4-12 emphasizes the balance between optimizing resource utilization and achieving the necessary precision. As the size of the link decreases, the system becomes more efficient in terms of resource usage. However, it also necessitates a higher error, potentially affecting the overall performance and accuracy of the system. The figure presents a valuable perspective on system design considerations in the constrained information exchange message passing algorithm in high dimensional data.

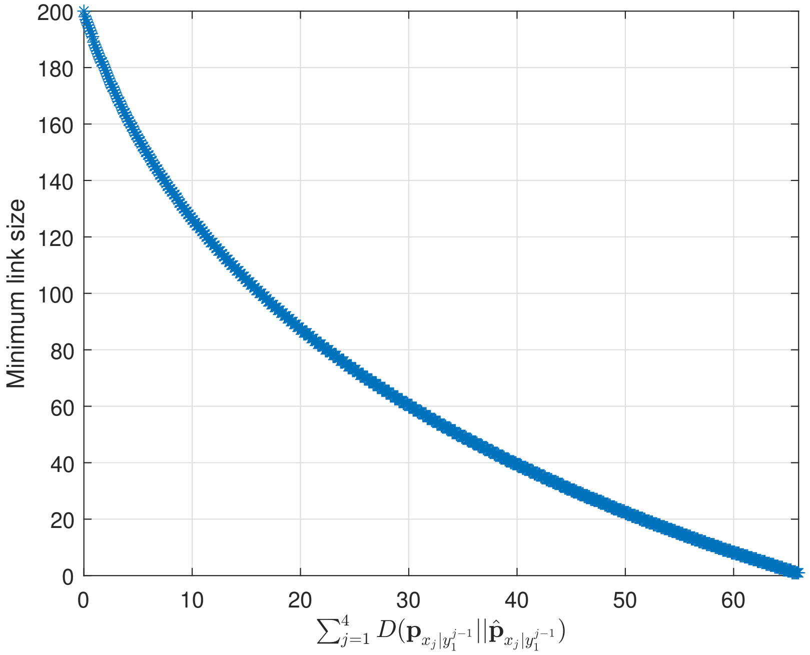


Figure 4-12: Error Threshold Impact on Minimum Link Size

Chapter 5

Conclusion

5.1 Summary of Work

In conclusion, this thesis has contributed to the advancement of algorithms for high-dimensional data in probabilistic graphical models by proposing and analyzing a constrained information exchange message passing algorithm that leverages insufficient statistics as messages.

Chapter 2 provided a comprehensive overview of the hidden Markov model and introduced the modified sum product algorithm. This algorithm, formulated in terms of the information vector and DTM/CDM matrix, enables the identification of critical information fragments for specific inference tasks. By identifying these important pieces of information, informed decisions can be made about which parts of the data to prioritize and which parts can be safely dropped or simplified.

Building upon the foundations established in Chapter 2, Chapter 3 introduced the constrained information exchange message passing algorithm. Chapter 3 introduced the index-specific posterior inference using constrained information exchange message passing algorithm. This sub-category of posterior inference provides a critical examination of specific scenarios where the interest lies either in a singular posterior inference at a specific index (referred to as single index-specific posterior inference) or in multiple posterior inferences at specific indexes (referred to as multiple index-specific posterior inference).

Chapter 4 introduced the index-free posterior inference using constrained information exchange message passing algorithm. Unlike the index-specific approach, where nodes may have different compression matrices, the index-free approach ensures that all nodes employ identical compression matrices. This sub-category of posterior inference provides a critical examination of specific scenarios where the interest lies either in a single step posterior inference (referred to as single step index-free posterior inference) or in multi-step posterior inferences (referred to as multi-step index-free posterior inference).

Notably, the algorithm sheds light on the trade-off between algorithm performance and resource utilization in modern data-driven applications using synthetic data. By reducing computational and communication complexity in various scenarios, it provides a practical and valuable solution for high-dimensional data problems, particularly when computational and communication resources are scarce.

In summary, this thesis has contributed to the field of probabilistic graphical models through the proposal and analysis of the constrained information exchange message passing algorithm. The algorithm's ability to identify critical information fragments and address the limitations of the unconstrained approach offers enhanced scalability, practicality, and usefulness in handling high-dimensional data. These advancements have the potential to drive advancements in data-driven research and decision-making across various domains.

5.2 Future Work

5.2.1 Constrained Information Exchange Message Passing Algorithm in Non-Linear Dynamical Systems

The constrained information exchange message passing algorithm has demonstrated success in optimizing computational and communication complexity in hidden Markov model and linear dynamical systems. This algorithm effectively identifies and transmits important fragments of information, leading to improved efficiency.

However, there remains a significant opportunity to explore the implementation of the constrained information exchange message passing algorithm in the context of non-linear dynamical systems. Non-linear systems are prevalent in numerous fields, including engineering, biology, physics, mathematics, and more. Therefore, conducting an in-depth analysis of this algorithm's performance in non-linear dynamical systems holds great significance for advancing research and applications in these domains.

By applying the constrained information exchange message passing algorithm to non-linear dynamical systems, we aim to enhance our understanding of the complex interactions and behaviors that arise in such systems. This exploration will enable us to develop more efficient models, improve predictions, and optimize decision-making processes in a wide range of practical scenarios.

Furthermore, the implications of successfully implementing this algorithm in non-linear dynamical systems are substantial. It can potentially revolutionize various fields by offering novel insights into intricate phenomena, aiding in the development of more efficient systems, and facilitating advancements in diverse areas of science and engineering.

5.2.2 Constrained Information Exchange Message Passing Algorithm in Loopy Graphs

Real-world problems often involve complex structures represented by loopy graphs. Effectively addressing these problems requires the development of efficient inference algorithms that can handle the inherent challenges posed by loops. To overcome these challenges, approximate inference algorithm is used.

In the future work, our objective is to analyze and design an innovative approximate inference algorithm specifically tailored for loopy graphs. By imposing constraints on the information exchange process, we aim to improve the efficiency of the inference procedure.

The utilization of constrained information exchange offers several advantages. By selectively transmitting critical fragments of information between nodes, we can mitigate the computational and communication complexities associated with loopy graphs. This approach allows us to strike a balance between accuracy and computational efficiency, making it particularly valuable for tackling large-scale, real-world problems.

To accomplish our goals, we will employ advanced techniques from the field of graphical models, inference algorithms, and information theory. By leveraging these methodologies, we can optimize the information exchange process and overcome the challenges presented by loopy graph structures.

Through extensive experimentation and analysis, we will evaluate the performance of our proposed algorithm against benchmark datasets and established inference methods. We will assess its accuracy, computational efficiency, and scalability in handling different types and sizes of loopy graphs. Additionally, we will investigate the impact of various constraints on the overall inference performance.

The findings from this research will contribute to the field of approximate inference in graphical models, particularly in the context of loopy graphs. By uncovering effective strategies for constrained information exchange, we can enhance our ability to solve complex real-world problems. The implications of this study extend to diverse domains such as computer vision, natural language processing, bioinformatics, and social network analysis, where loopy graph structures are prevalent.

5.2.3 Constrained Information Exchange Viterbi Algorithm

The Viterbi algorithm holds immense significance in probabilistic graphical models as it enables the identification of the most probable configuration. However, to further enhance the algorithm's capabilities, we propose exploring the concept of constrained information exchange within the Viterbi framework.

In our future work, we will investigate the potential implications and benefits of incorporating constrained information exchange into the Viterbi algorithm. By selectively exchanging crucial fragments of information, we aim to improve the efficiency and applicability of the algorithm in the case of high-dimensional data.

The introduction of constraints in the information exchange process can offer several advantages. By focusing on transmitting only the most relevant information between nodes, we can significantly reduce computational and communication complexity and improve overall runtime efficiency.

Our research will involve the development and analysis of the constrained information exchange Viterbi algorithm. We will explore various techniques and methodologies to effectively integrate the constraints while preserving the core functionality of the original algorithm. Additionally, we will evaluate the performance of the constrained information exchange Viterbi algorithm across different probabilistic graphical models, assessing its impact on the quality of the final configuration.

The implications of this research extend to numerous domains where probabilistic graphical models are widely employed, such as natural language processing, speech recognition, and bioinformatics. By enhancing the Viterbi algorithm through constrained information exchange, we can improve the efficiency of model predictions, enhance decision-making processes, and advance research in these fields.

Appendix A

Mathematical Proofs

A.1 Proof of Lemma 2.4

$$
\begin{aligned}
\log\left(\frac{p_1(x)}{p_0(x)}\right) &= \log\left(1+\epsilon\frac{\phi_1(x)}{\sqrt{p_o(x)}}\right) \\
&\text{using second order taylor series} \\
&= \epsilon\frac{\phi_1(x)}{\sqrt{p_0(x)}} - \frac{\epsilon^2}{2}\frac{\phi_1^2(x)}{p_o(x)} + o(\epsilon^2)
\end{aligned}
$$

$$
\begin{aligned}
&D(\mathbf{p}_1||\mathbf{p}_2) \\
&= \sum_{\mathsf{x}\in\mathcal{X}} p_1(\mathsf{x})\log\frac{p_1(\mathsf{x})}{p_2(\mathsf{x})} \\
&= \sum_{\mathsf{x}\in\mathcal{X}} p_0(\mathsf{x})\log\frac{p_1(\mathsf{x})}{p_2(\mathsf{x})} + \sum_{\mathsf{x}\in\mathcal{X}}(p_1(\mathsf{x})-p_0(\mathsf{x}))\log\frac{p_1(\mathsf{x})}{p_2(\mathsf{x})} \\
&= \sum_{\mathsf{x}\in\mathcal{X}} p_0(\mathsf{x})\left(\log\frac{p_1(\mathsf{x})}{p_0(\mathsf{x})}-\log\frac{p_2(\mathsf{x})}{p_0(\mathsf{x})}\right) + \sum_{\mathsf{x}\in\mathcal{X}}(p_1(\mathsf{x})-p_0(\mathsf{x}))\left(\log\frac{p_1(\mathsf{x})}{p_0(\mathsf{x})}-\log\frac{p_2(\mathsf{x})}{p_0(\mathsf{x})}\right) \\
&= \sum_{\mathsf{x}\in\mathcal{X}} p_0(\mathsf{x})\left(\log\frac{p_1(\mathsf{x})}{p_0(\mathsf{x})}-\log\frac{p_2(\mathsf{x})}{p_0(\mathsf{x})}\right) + \sum_{\mathsf{x}\in\mathcal{X}}(\epsilon\phi_1(x)\sqrt{p_0(x)})\left(\log\frac{p_1(\mathsf{x})}{p_0(\mathsf{x})}-\log\frac{p_2(\mathsf{x})}{p_0(\mathsf{x})}\right) \\
&= \epsilon\sum_{\mathsf{x}\in\mathcal{X}} p_0(\mathsf{x})\left(\frac{\phi_1(x)-\phi_2(x)}{\sqrt{p_0(x)}}\right) - \frac{\epsilon^2}{2}\sum_{\mathsf{x}\in\mathcal{X}} p_0(\mathsf{x})\left(\frac{\phi_1^2(x)-\phi_2^2(x)}{p_0(x)}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ \epsilon^2 \sum_{\mathsf{x}\in\mathcal{X}} (\phi_1(x)\sqrt{p_0(x)}) \left(\frac{\phi_1(x) - \phi_2(x)}{\sqrt{p_0(x)}}\right) - \frac{\epsilon^3}{2} \sum_{\mathsf{x}\in\mathcal{X}} (\phi_1(x)\sqrt{p_0(x)}) \left(\frac{\phi_1^2(x) - \phi_2^2(x)}{p_0(x)}\right) + o(\epsilon^2) \\
&= 0 - \frac{\epsilon^2}{2} \sum_{\mathsf{x}\in\mathcal{X}} (\phi_1^2(x) - \phi_2^2(x)) + \epsilon^2 \sum_{\mathsf{x}\in\mathcal{X}} \phi_1(x)\left(\phi_1(x) - \phi_2(x)\right) + o(\epsilon^2) \\
&= \frac{\epsilon^2}{2}(||\boldsymbol{\phi}_2||^2 - ||\boldsymbol{\phi}_1||^2 + 2||\boldsymbol{\phi}_1||^2 - 2\langle \boldsymbol{\phi}_1, \boldsymbol{\phi}_2\rangle) + o(\epsilon^2) \\
&= \frac{\epsilon^2}{2}||\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2||^2 + o(\epsilon^2)
\end{aligned}
$$

A.2 Proof of Lemma 2.7

$$
\begin{aligned}
p_{\mathsf{x}_i|\mathsf{y}_1^i}(x_i|\check{y}_1^i) &= \frac{p_{\mathsf{x}_i}(x_i)}{p_{\mathsf{y}_1^i}(\check{y}_1^i)} \frac{p_{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(x_i|\check{y}_1^i) p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i) p_{\mathsf{y}_1^{i-1}}(\check{y}_1^{i-1}) p_{\mathsf{y}_i}(\check{y}_i)}{p_{\mathsf{x}_i}(x_i)^2} \\
&= \underbrace{p_{\mathsf{x}_i}(x_i)\frac{p_{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(x_i|\check{y}_1^i) p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i)}{p_{\mathsf{x}_i}(x_i)^2}}_{a} \underbrace{\frac{p_{\mathsf{y}_1^{i-1}}(\check{y}_1^{i-1}) p_{\mathsf{y}_i}(\check{y}_i)}{p_{\mathsf{y}_1^i}(\check{y}_1^i)}}_{b^{-1}}
\end{aligned}
$$

$$
\begin{aligned}
a &= p_{\mathsf{x}_i}(x_i)\frac{p_{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(x_i|\check{y}_1^i) p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i)}{p_{\mathsf{x}_i}(x_i)^2} \\
&= p_{\mathsf{x}_i}(x_i)\left(1 + \frac{\epsilon\phi^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})}(x_i|\check{y}_1^{i-1})}{\sqrt{p_{\mathsf{x}_i}(x_i)}}\right)\left(1 + \frac{\epsilon\phi^{(\mathsf{x}_i|\mathsf{y}_i)}(x_i|\check{y}_i)}{\sqrt{p_{\mathsf{x}_i}(x_i)}}\right) \\
&= p_{\mathsf{x}_i}(x_i) + \epsilon\sqrt{p_{\mathsf{x}_i}(x_i)}\left[\phi^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})}(x_i|\check{y}_1^{i-1}) + \phi^{(\mathsf{x}_i|\mathsf{y}_i)}(x_i|\check{y}_i)\right] + o(\epsilon)
\end{aligned}
$$

$$
\begin{aligned}
b &= \frac{p_{\mathsf{y}_1^{i-1}}(\check{y}_1^{i-1}) p_{\mathsf{y}_i}(\check{y}_i)}{p_{\mathsf{y}_1^i}(\check{y}_1^i)} \\
&= \sum_{x_i\in\mathcal{X}} p_{\mathsf{x}_i}(x_i)\frac{p_{\mathsf{x}_i|\mathsf{y}_1^{i-1}}(x_i|\check{y}_1^i) p_{\mathsf{x}_i|\mathsf{y}_i}(x_i|\check{y}_i)}{p_{\mathsf{x}_i}(x_i)^2} \\
&= \sum_{x_i\in\mathcal{X}} p_{\mathsf{x}_i}(x_i) + \epsilon\sqrt{p_{\mathsf{x}_i}(x_i)}\left[\phi^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})}(x_i|\check{y}_1^{i-1}) + \phi^{(\mathsf{x}_i|\mathsf{y}_i)}(x_i|\check{y}_i)\right] + o(\epsilon) \\
&= 1 + o(\epsilon)
\end{aligned}
$$

Combining a and b

$$
\begin{aligned}
p_{\mathsf{x}_i|\mathsf{y}_1^i}(x_i|\check{y}_1^i) &= p_{\mathsf{x}_i}(x_i) + \epsilon\sqrt{p_{\mathsf{x}_i}(x_i)}\left[\phi^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})}(x_i|\check{y}_1^{i-1}) + \phi^{(\mathsf{x}_i|\mathsf{y}_i)}(x_i|\check{y}_i)\right] + o(\epsilon) \\
\frac{p_{\mathsf{x}_i|\mathsf{y}_1^i}(x_i|\check{y}_1^i) - p_{\mathsf{x}_i}(x_i)}{\epsilon\sqrt{p_{\mathsf{x}_i}(x_i)}} &= \phi^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})}(x_i|\check{y}_1^{i-1}) + \phi^{(\mathsf{x}_i|\mathsf{y}_i)}(x_i|\check{y}_i) + o(\epsilon)
\end{aligned}
$$

$$\implies \boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_1^i)}(.|\check{y}_1^i) = \boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})}(.|\check{y}_1^{i-1}) + \boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_i)}(.|\check{y}_i) + o(\epsilon)$$

$$\begin{aligned}
\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_1^{i-1})}(.|\check{y}_1^{i-1}) &= \mathbf{B}_{\mathsf{x}_{i-1},\mathsf{x}_i}\boldsymbol{\phi}^{(\mathsf{x}_{i-1}|\mathsf{y}_1^{i-1})}(.|\check{y}_1^{i-1}) \\
&= \mathbf{B}_{\mathsf{x}_{i-1},\mathsf{x}_i}\boldsymbol{\phi}^{(\mathsf{x}_{i-1}|\mathsf{y}_1^{i-2})}(.|\check{y}_1^{i-2}) + \mathbf{B}_{\mathsf{x}_{i-1},\mathsf{x}_i}\boldsymbol{\phi}^{(\mathsf{x}_{i-1}|\mathsf{y}_{i-1})}(.|\check{y}_{i-1}) + o(\epsilon)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_1^i)}(.|\check{y}_1^i) &= \sum_{j=1}^{i}\left(\prod_{k=j}^{i-1}\mathbf{B}_{\mathsf{x}_k,\mathsf{x}_{k+1}}\right)\boldsymbol{\phi}^{(\mathsf{x}_j|\mathsf{y}_j)}(.|\check{y}_j) + o(\epsilon) \\
&= \sum_{j=1}^{i}\boldsymbol{\phi}^{(\mathsf{x}_i|\mathsf{y}_j)}(.|\check{y}_j) + o(\epsilon)
\end{aligned}$$

Appendix B

Multiple Index-Specific Posterior Inference Using CDM

Given the singularity of the CDM matrix, it is crucial to introduce perturbations in order to transform it into a non-singular matrix. This perturbation process involves making small adjustments to the elements of the matrix to ensure a non-zero determinant. By perturbing the CDM matrix, we can guarantee that it attains full rank and overcomes the singularity issue. These perturbations are essential for enabling the application of relevant theorems, for the multiple index-specifc posterior inference.

$$\tilde{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}} = \hat{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}} + \epsilon_1 \mathbf{I} \tag{B.1}$$

The Frobenius norm of the perturbed matrix

$$\begin{aligned} \|\tilde{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}}\|_F^2 &= \|\hat{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}}\|_F^2 + 2\epsilon_1 \text{trace}(\hat{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}}) + \epsilon_1^2 |\mathcal{X}| && \text{(B.2)} \\ &\approx \|\hat{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}}\|_F^2 && \text{(B.3)} \end{aligned}$$

Theorem B.1. *Let* $\tilde{\mathbf{B}}_{\mathsf{x}_j,\mathsf{x}_{j+1}} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$*,* $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$*,*$\mathbf{G}_j \in \mathbb{R}^{|\mathcal{X}|\times r}$*, and* $\mathbf{H}_j \in \mathbb{R}^{r\times|\mathcal{X}|}$

where $j \in \{1, \cdots, i\}$*. The SVD of*

$$\left\| \begin{bmatrix} \sqrt{\gamma}(\left(\prod_{k=j}^{i-1} \tilde{\mathbf{B}}_{\mathsf{x}_k,\mathsf{x}_{k+1}}\right)) \\ \sqrt{(1-\gamma)}\left(\prod_{k=j}^{i} \tilde{\mathbf{B}}_{\mathsf{x}_k,\mathsf{x}_{k+1}}\right) \end{bmatrix} \right\|_F^2 = \begin{bmatrix} \underbrace{\mathbf{U}_{j,1}}_{|\mathcal{X}|\times|\mathcal{X}|} & \\ & \underbrace{\mathbf{G}}_{2|\mathcal{X}|\times|\mathcal{X}|} \\ \underbrace{\mathbf{U}_{j,2}}_{|\mathcal{X}|\times|\mathcal{X}|} & \end{bmatrix} \begin{bmatrix} \underbrace{\overline{\mathbf{S}}_j}_{|\mathcal{X}|\times|\mathcal{X}|} \\ \mathbf{0}_{|\mathcal{X}|\times|\mathcal{X}|} \end{bmatrix} \underbrace{\mathbf{V}_j^T}_{|\mathcal{X}|\times|\mathcal{X}|}$$

Therefore, the optimal matrices are

$$\begin{aligned} \mathbf{H}_j &= \overline{\mathbf{S}}_j^{(r)} \mathbf{V}_j^{(r)T} \\ \mathbf{G}_j &= \frac{1}{\sqrt{\gamma}} (\tilde{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}}^{-1})^{i-1-j} \mathbf{U}_{j,1} \begin{bmatrix} \mathbf{I}_{r\times r} \\ 0_{|\mathcal{X}|-r\times r} \end{bmatrix} \\ &= \frac{1}{\sqrt{\gamma}} (\tilde{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}})^{j+1-i} \mathbf{U}_{j,1} \begin{bmatrix} \mathbf{I}_{r\times r} \\ 0_{|\mathcal{X}|-r\times r} \end{bmatrix} \\ \mathbf{F} &= \tilde{\mathbf{B}}_{\mathsf{x}_i,\mathsf{x}_{i+1}} \end{aligned}$$

where $\overline{\mathbf{S}}_j^{(r)}$ *is* $r \times r$ *matrix with* r *largest singular values,* $\mathbf{V}_j^{(r)}$ *is* $|\mathcal{X}| \times r$ *matrix is the corresponding right singular vectors.*

Theorem B.2. *Let* $\tilde{\mathbf{B}} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$*,* $\mathbf{F} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|}$*,*$\mathbf{G}_j \in \mathbb{R}^{|\mathcal{X}|\times r}$*, and* $\mathbf{H}_j \in \mathbb{R}^{r\times|\mathcal{X}|}$ *where* $j \in \{1, \cdots, i\}$*. Without loss of generality, the SVD of the concatenated matrix at* $j = 1$

$$\left\| \begin{matrix} \sqrt{\gamma_1}\tilde{\mathbf{B}}^i \\ \sqrt{\gamma_2}\tilde{\mathbf{B}}^{i-1} \\ \vdots \\ \sqrt{\gamma_i}\tilde{\mathbf{B}} \end{matrix} \right\|_F^2 = \begin{bmatrix} \underbrace{\mathbf{U}_{1,1}}_{|\mathcal{X}|\times|\mathcal{X}|} & \\ \underbrace{\mathbf{U}_{1,2}}_{|\mathcal{X}|\times|\mathcal{X}|} & \underbrace{\mathbf{G}}_{(i)|\mathcal{X}|\times|\mathcal{X}|} \\ \vdots & \\ \underbrace{\mathbf{U}_{1,i}}_{|\mathcal{X}|\times|\mathcal{X}|} & \end{bmatrix} \begin{bmatrix} \underbrace{\overline{\mathbf{S}}_1}_{|\mathcal{X}|\times|\mathcal{X}|} \\ \mathbf{0}_{(i-1)|\mathcal{X}|\times|\mathcal{X}|} \end{bmatrix} \underbrace{\mathbf{V}_1^T}_{|\mathcal{X}|\times|\mathcal{X}|}$$

where $\mathbf{U}_{1,t}$ *and* $\mathbf{V}_1$ *are the left and right singular vectors, and* $\overline{\mathbf{S}}_1$ *are the singular* *values. The optimal matrices for*

$$\left\| \begin{array}{c} \sqrt{\gamma_1}(\tilde{\mathbf{B}}^i - \mathbf{F}^{i-1}\mathbf{G}_1\mathbf{H}_1) \\ \sqrt{\gamma_2}(\tilde{\mathbf{B}}^{i-1} - \mathbf{F}^{i-2}\mathbf{G}_1\mathbf{H}_1) \\ \vdots \\ \sqrt{\gamma_i}(\tilde{\mathbf{B}} - \mathbf{G}_1\mathbf{H}_1) \end{array} \right\|_F^2$$

are

$$\mathbf{H}_1 = \overline{\mathbf{S}}_1^{(r)}\mathbf{V}_1^{(r)T} \ ; \ \mathbf{G}_1 = \frac{1}{\sqrt{\gamma_i}}\mathbf{U}_{1,i}\left[\begin{array}{c} \mathbf{I}_{r\times r} \\ 0_{|\mathcal{X}|-r\times r} \end{array} \right] \ ; \ \mathbf{F} = \tilde{\mathbf{B}}$$

where $\overline{\mathbf{S}}_1^{(r)}$ *is* $r \times r$ *matrix with* r *largest singular values, and* $\mathbf{V}_1^{(r)}$ *is* $|\mathcal{X}| \times r$ *matrix with the corresponding right singular vectors. The same will be done for the other concatenated matrices for different* j *values and the optimal matrices are*

$$\mathbf{H}_j = \overline{\mathbf{S}}_j^{(r)}\mathbf{V}_j^{(r)T} \ ; \ \mathbf{G}_j = \frac{1}{\sqrt{\gamma_{i-j+1}}}\mathbf{U}_{j,i-j+1}\left[\begin{array}{c} \mathbf{I}_{r\times r} \\ 0_{|\mathcal{X}|-r\times r} \end{array} \right]$$

Bibliography

[1] L. E. Sucar, “Probabilistic graphical models,” *Advances in Computer Vision and Pattern Recognition. London: Springer London. doi*, vol. 10, no. 978, p. 1, 2015.

[2] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

[3] A. Khalifeh, K. A. Darabkh, A. M. Khasawneh, I. Alqaisieh, M. Salameh, A. Al-Abdala, S. Alrubaye, A. Alassaf, S. Al-HajAli, R. Al-Wardat *et al.*, “Wireless sensor networks for smart cities: Network design, implementation and performance evaluation,” *Electronics*, vol. 10, no. 2, p. 218, 2021.

[4] L. M. Oliveira and J. J. Rodrigues, “Wireless sensor networks: A survey on environmental monitoring.” *J. Commun.*, vol. 6, no. 2, pp. 143–151, 2011.

[5] S.-H. Choi, B.-K. Kim, J. Park, C.-H. Kang, and D.-S. Eom, “An implementation of wireless sensor network for security system using bluetooth,” *IEEE Transactions on Consumer Electronics*, vol. 50, no. 1, pp. 236–244, 2004.

[6] M. Majid, S. Habib, A. R. Javed, M. Rizwan, G. Srivastava, T. R. Gadekallu, and J. C.-W. Lin, “Applications of wireless sensor networks and internet of things frameworks in the industry revolution 4.0: A systematic literature review,” *Sensors*, vol. 22, no. 6, p. 2087, 2022.

[7] P. Gope, A. K. Das, N. Kumar, and Y. Cheng, “Lightweight and physically secure anonymous mutual authentication protocol for real-time data access in industrial

wireless sensor networks," *IEEE transactions on industrial informatics*, vol. 15, no. 9, pp. 4957–4968, 2019.

[8] J. R. Norris, *Markov chains.* Cambridge university press, 1998, no. 2.

[9] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning.* Springer, 2006, vol. 4, no. 4.

[10] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[11] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition.* Prentice-Hall, Inc., 1993.

[12] F. Jelinek, *Statistical methods for speech recognition.* MIT press, 1998.

[13] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 179–190, 1983.

[14] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, ""your word is my command": Google search by voice: A case study," *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, pp. 61–90, 2010.

[15] D. Jurafsky and J. Martin, *Speech and Language Processing, 2nd Edition*, 2nd ed. Upper Saddle River, N.J: Prentice Hall, May 2008.

[16] A. Ekbal, S. Mondal, and S. Bandyopadhyay, "Pos tagging using hmm and rule-based chunking," *The Proceedings of SPSAL*, vol. 8, no. 1, pp. 25–28, 2007.

[17] M. Taulé, M. A. Martí, and M. Recasens, "Ancora: Multilevel annotated corpora for catalan and spanish." in *Lrec*, 2008.

[18] M. A. M. Antonín, M. T. Delor, L. Màrquez, and M. Bertran, "Anotación semi-automática con papeles temáticos de los corpus cess-ece," *Procesamiento del Lenguaje Natural*, no. 38, pp. 67–76, 2007.

[19] G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in *Proceedings of the 40th annual meeting of the association for computational linguistics*, 2002, pp. 473–480.

[20] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge university press, 1998.

[21] S. R. Eddy, "Hidden markov models," *Current opinion in structural biology*, vol. 6, no. 3, pp. 361–365, 1996.

[22] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden markov models in computational biology: Applications to protein modeling," *Journal of molecular biology*, vol. 235, no. 5, pp. 1501–1531, 1994.

[23] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic dna," *Journal of molecular biology*, vol. 268, no. 1, pp. 78–94, 1997.

[24] B.-J. Yoon, "Hidden markov models and their applications in biological sequence analysis," *Current genomics*, vol. 10, no. 6, pp. 402–415, 2009.

[25] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer *et al.*, "The pfam protein families database," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D138–D141, 2004.

[26] R. D. Finn, J. Clements, and S. R. Eddy, "Hmmer web server: interactive sequence similarity searching," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W29–W37, 2011.

[27] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature methods*, vol. 5, no. 1, pp. 16–18, 2008.

[28] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate (corresp.)," *IEEE Transactions on information theory*, vol. 20, no. 2, pp. 284–287, 1974.

[29] J. G. Proakis, *Digital signal processing: principles, algorithms, and applications, 4/E.* Pearson Education India, 2007.

[30] B. Fritchman, “A binary channel characterization using partitioned markov chains,” *IEEE transactions on Information Theory*, vol. 13, no. 2, pp. 221–227, 1967.

[31] F. Chen and S. Kwong, “Multiuser detection using hidden markov model,” *IEEE Transactions on Vehicular Technology*, vol. 58, no. 1, pp. 107–115, 2008.

[32] J. Hagenauer and P. Hoeher, “A viterbi algorithm with soft-decision outputs and its applications,” in *1989 IEEE Global Telecommunications Conference and Exhibition'Communications Technology for the 1990s and Beyond'.* IEEE, 1989, pp. 1680–1686.

[33] G. D. Forney, “The viterbi algorithm,” *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[34] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “A survey on sensor networks,” *IEEE Communications magazine*, vol. 40, no. 8, pp. 102–114, 2002.

[35] J. Yick, B. Mukherjee, and D. Ghosal, “Wireless sensor network survey,” *Computer networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[36] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, “Wireless sensor networks for habitat monitoring,” in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, 2002, pp. 88–97.

[37] N. Wang, N. Zhang, and M. Wang, “Wireless sensors in agriculture and food industry—recent development and future perspective,” *Computers and electronics in agriculture*, vol. 50, no. 1, pp. 1–14, 2006.

[38] L. Yu, N. Wang, and X. Meng, “Real-time forest fire detection with wireless sensor networks,” in *Proceedings. 2005 International Conference on Wireless Com-*

munications, Networking and Mobile Computing, 2005., vol. 2. Ieee, 2005, pp. 1214–1217.

[39] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: A survey," *Ad hoc networks*, vol. 7, no. 3, pp. 537–568, 2009.

[40] S.-L. Huang, A. Makur, G. W. Wornell, and L. Zheng, "On universal features for high-dimensional learning and inference," *arXiv preprint arXiv:1911.09105*, 2019.

[41] T. M. Cover, *Elements of information theory.* John Wiley & Sons, 1999.

[42] S. Borade and L. Zheng, "Euclidean information theory," in *2008 IEEE International Zurich Seminar on Communications.* IEEE, 2008, pp. 14–17.

[43] E. Abbe and L. Zheng, "Linear universal decoding for compound channels," *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 5999–6013, 2010.

[44] R. G. Gallager, *Information theory and reliable communication.* Springer, 1968, vol. 588.

[45] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[46] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," 1993.

[47] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[48] L. Y. Kolotilina, "The strengthened versions of the additive and multiplicative weyl inequalities," *Journal of Mathematical Sciences*, vol. 127, no. 3, pp. 1976–1987, 2005.

[49] Y. Collette and P. Siarry, *Multiobjective optimization: principles and case studies.* Springer Science & Business Media, 2004.