



THESE DE DOCTORAT DE SORBONNE UNIVERSITE

Spécialité
Electroniques et Communications

Ecole doctorale
EDITE ED 130

Présentée par
Zilu Zhao

Pour obtenir le grade de
DOCTEUR de SORBONNE UNIVERSITE

Sujet de la thèse :
**Bayesian Learning with Applications to Cell-Free
Massive MIMO**

soutenue le **26.06.2026**

devant le jury composé de :

Prof. Dirk Slock	EURECOM, FR	Directeur de thèse
Dr. Maxime Guillaud	INRIA, FR	Rapporteur/ Examineur
Prof. Charly Poulliat	INP-ENSEEIH, FR	Rapporteur/ Examineur
Prof. Laura Cottatellucci	FAU Erlangen-Nürnberg, DE	Examinatrice
Prof. Wolfgang Utschick	TU Munich, DE	Examineur
Prof. Roberto Morabito	EURECOM, FR	Examineur

Contents

Abstract	7
List of Figures	9
1 Introduction	10
1.1 Prior Works	11
1.1.1 Useful Bayesian Inference Tools	11
1.1.2 Prior Works for Linear Estimation	13
1.1.3 Prior Works for Bilinear Estimation	13
1.2 Main Contributions	14
1.2.1 Main Contributions in Linear Estimations	14
1.2.2 Main Contributions in Bilinear Estimations	15
2 Detailed Introduction to BFE and Message Passing Algorithms	17
2.1 Bethe Free Energy	17
2.1.1 Tree Structured Factorization Example	17
2.1.2 Analysis of Tree Structured Factorization	18
2.1.3 Bethe Free Energy in Non-Cyclic Graph	23
2.1.4 Bethe Free Energy Extended to Graph with Cycles	23
2.2 Relation Between Bethe Free Energy and Message Passing Algorithms . . .	23
2.2.1 Relation to Belief Propagation (Strict Marginal Consistency Con-	
straints)	25
2.2.2 Relation to Expectation Propagation (Relaxed Moment Consistency	
Constraints)	26
I Linear Estimation	28
3 Vector Approximate Message Passing for Not So Large N.I.I.D. Gener-	29
alized I/O Linear Models	
3.1 Introduction	29
3.1.1 Prior Work	29
3.1.2 Main Contribution	30
3.2 Derivation under EP Framework	30
3.2.1 Output Node	31
3.2.2 Input Node	33
3.3 Simulation Results	35
3.4 Concluding Remarks	36

4	Bethe Free Energy and Extrinsic in Approximate Message Passing	37
4.1	Introduction	37
4.1.1	Prior Work	37
4.1.2	Main Contributions	38
4.2	Bethe Free Energy of Generalized Linear Model	38
4.2.1	Bethe Free Energy	38
4.2.2	BFE of GLM	39
4.3	Relation to Message Passing and its Stationary Points	40
4.3.1	Approximation of Prior $p(\mathbf{x})$	41
4.3.2	Extrinsic for Output Node $\mathbf{z}$	41
4.3.3	Approximation of Likelihood $p(\mathbf{y} \mathbf{z})$	43
4.4	Derivation of LSL-BFE	43
4.5	Relation to CWCU MMSE Estimator	44
4.6	Concluding Remarks	46
5	Towards Convergent Approximate Message Passing by Alternating Con- strained Minimization of Bethe Free Energy	47
5.1	Introduction	47
5.1.1	Contributions	47
5.2	System Model	48
5.3	Proposed AMBGAMP	48
5.3.1	Update of $\mathbf{u}$	50
5.3.2	Update of $q_{\mathbf{x}}$	51
5.3.3	Update of $\{q_{\mathbf{z}}\}$	51
5.3.4	Update of $\{\mathbf{s}\}$ (ADMM style)	52
5.3.5	Update of $\{\boldsymbol{\tau}_p, \boldsymbol{\tau}_s\}$	52
5.4	Convergence to LMMSE	53
5.5	Convergence of the Variance Subsystem	54
5.6	Large System Analysis with n.i.i.d. $\mathbf{A}$	55
5.7	Convergence of the Mean Subsystem	56
5.8	Simulation Results	57
5.9	Concluding Remarks	58
II	Bilinear Estimation	59
6	Bilinear Hybrid Expectation Maximization and Expectation Propaga- tion for Semi-Blind Channel Estimation	60
6.1	Introduction	60
6.1.1	Prior Works	60
6.1.2	Main Contribution	61
6.2	System Model	61
6.2.1	Orthogonal Pilot	62
6.3	EP Approximation	62
6.4	EM-EP	62
6.5	Loop Free Expectation Propagation	66
6.6	Simulation Results	68
6.7	Conclusions	69

7	Decentralized Expectation Propagation for Semi-Blind Channel Estimation in Cell-Free Networks	71
7.1	Introduction	71
7.1.1	Main Contributions	72
7.2	System Model	72
7.2.1	Orthogonal Pilot sequences	72
7.2.2	Expectation Propagation on Semi-Blind structure	72
7.3	Message Passing Derivations	73
7.3.1	Message from $\Psi_{2,lt}$ to x_{kt}	73
7.3.2	Message from $\Psi_{2,lt}$ to $\mathbf{h}_{lk}$	74
7.3.3	Message form $\Psi_{3,lg}$ to $\mathbf{h}_{lk}$	75
7.4	Asymptotic Behaviors in Large Systems	76
7.5	Simplification of the Messages	78
7.6	Decentralized Method	78
7.7	Simulation Results	79
7.8	Conclusions	80
8	Decentralized Message-Passing for Semi-Blind Channel Estimation in Cell-Free Systems Based on Bethe Free Energy Optimization	81
8.1	Introduction	81
8.1.1	Orthogonal Pilots	81
8.1.2	Factored Joint Distribution	82
8.1.3	Notations	82
8.1.4	Prior Work	82
8.1.5	Main Contributions	83
8.2	Bethe Free Energy Optimization Framework	83
8.2.1	Bethe Approximation with Constraints	83
8.2.2	Bethe Free Energy Optimization	84
8.3	Detailed Derivations	85
8.3.1	Update of $\mu_{f_{z_l}; \mathbf{z}_{lk}}$	86
8.3.2	Update of $\mu_{f_{\mathbf{h}_{lG_g}}; \mathbf{h}_{lk}}$	86
8.3.3	Update of $\mu_{f_{\mathbf{x}_k}; \mathbf{x}_k}$	86
8.3.4	Update of $\mu_{\delta_{lk}; \mathbf{x}_k}$	86
8.3.5	Update of $\mu_{\delta_{lk}; \mathbf{h}_{lk}}$	87
8.3.6	Update of $\mu_{\delta_{lk}; \mathbf{z}_{lk}}$	87
8.4	Decentralization Method	88
8.5	Simulation Results	89
8.6	Conclusions	90
9	Hierarchical Expectation Propagation for Semi-Blind Channel Estimation in Cell-Free Networks	91
9.1	Introduction	91
9.1.1	Continuous Input	91
9.1.2	Discrete Input	92
9.1.3	Main Contributions	92
9.2	System Model	92
9.2.1	Orthogonal Pilot sequences	92
9.2.2	Hierarchical Data Symbol	93

9.2.3	Notations	93
9.3	Message Passing Derivations	94
9.3.1	Message from $\delta_{m,lt}$ to $x_{m,kt}$	95
9.3.2	Message from $\delta_{m,lt}$ to $\mathbf{h}_{lk}$	96
9.3.3	Message from $\delta_{m,lt}$ to $\mathbf{z}_{m,lt}$	96
9.3.4	Message from $f_{\mathbf{H}_{lG_g}}$ to $\mathbf{h}_{lk}$	97
9.3.5	Message from $f_{\mathbf{y}_{lt}}$ to $\mathbf{z}_{m,lt}$	97
9.4	Simulation Results	97
9.5	Conclusions	98
10	Distributed Iterative ML and Message Passing for Grant-Free Cell-Free Massive MIMO Systems	99
10.1	Introduction	99
10.1.1	Prior Works	99
10.1.2	Main Contributions	100
10.2	System Model	100
10.2.1	Orthogonal Pilot sequences	100
10.3	Iterative ML for User Activity Detection	101
10.4	Distributed VL-EP for Pseudo-Prior	103
10.4.1	Expectation Step	104
10.4.2	Maximization Step	105
10.5	Bilinear VB-EP	105
10.6	Iterative Solution to VB-EP	107
10.6.1	Update of $\mu_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}}$	107
10.6.2	Update of $\mu_{f_{\mathbf{H}_{lG_g}}; \mathbf{h}_{lk}}$	107
10.6.3	Update of $\mu_{\delta_{lk}; \mathbf{x}_k}$	108
10.6.4	Update of $\mu_{\delta_{lk}; \mathbf{h}_{lk}}$	108
10.6.5	Update of $\mu_{\delta_{lk}; \mathbf{z}_{lk}}$	108
10.7	Simulation Results	109
10.8	Conclusions	110
11	Approximating Univariate Factored Distributions via Message-Passing Algorithms	111
11.1	Introduction	111
11.1.1	Prior Works	111
11.1.2	Main Contribution	112
11.2	Problem Formulation	113
11.3	VDBP-Based Approximation for Univariate Factored PDFs	113
11.3.1	Outline of the VDBP	113
11.3.2	Derivation of the VDBP	116
11.3.3	Approximated Statistics	118
11.4	Preliminaries for Expectation Propagation	119
11.4.1	Factor Graph of the Given Problem in EP	120
11.5	Persistent EP	120
11.6	Analytic Continuation EP	123
11.6.1	GMM Case	126
11.7	Simulation Results	128
11.8	Conclusions	129

12	Avoiding Non-Integrable Beliefs in Expectation Propagation	131
12.1	Introduction	131
12.1.1	Main Contributions	131
12.2	Analytic Continuation Expectation Propagation (ACEP)	132
12.2.1	Constrained KLD Optimization in Gaussian Projection	132
12.3	EP with Checking (EPC)	134
12.4	System Model for Generalized Linear Model	134
12.5	Derivation for GLM	135
12.5.1	$f_{y_n} \rightarrow z_n$	135
12.5.2	$f_{x_k} \rightarrow x_k$	136
12.5.3	$f_z \rightarrow x_k$	136
12.5.4	$f_z \rightarrow z_n$	137
12.5.5	$z \rightarrow f_y$	138
12.5.6	$z \rightarrow f_z$	138
12.5.7	$x \rightarrow f_x$	139
12.5.8	$x \rightarrow f_z$	139
12.6	Useful Relations	140
12.7	Gaussian Mixture Model (GMM)	144
12.8	Simulation Results	145
12.9	Conclusions	146
13	Conclusions	147
13.1	Future Work	148

Abstract

In this thesis, message-passing-based algorithms for the uplink of Cell-Free Massive Multiple-Input Multiple-Output (CF-MaMIMO) communications are investigated. In CF networks, all user equipments (UEs) within a region are simultaneously served by all access points (APs). As a result, the number of users may exceed the length of pilot sequences, leading to pilot contamination. To address this issue, semi-blind methods are required, in which APs jointly estimate channel coefficients and UE data symbols. Pilot-based channel estimation can be formulated as a linear estimation problem, while data-based estimation leads to a bilinear estimation problem. Exact Bayesian inference involves high-dimensional integrals and is NP-hard. Therefore, iterative approximation methods, such as message-passing algorithms, are essential for practical CF communication systems. The close relationship between constrained Bethe Free Energy (BFE) optimization and message-passing algorithms is analyzed. Pilot-based channel estimation is generalized as a sparse signal recovery problem under a generalized linear model (GLM), which is a fundamental problem in signal processing. The approximate posterior, along with the corresponding equivalent Gaussian prior and noise, is obtained using Expectation Propagation (EP) on a loop-free factorization of the GLM. A constrained BFE formulation that shares the same stationary points as the corresponding EP algorithm is studied. Furthermore, an alternative derivation of the large-system-limit (LSL) approximation of the BFE is presented. Based on this LSL approximation, a provably convergent iterative method is proposed. In the second part, this thesis focuses on bilinear estimation problems. A hybrid method combining Expectation Maximization (EM) and EP (EM-EP) is proposed, where EP mitigates multi-user interference and EM converts the single-user bilinear estimation into two linear estimation problems. By employing a loop-free factorization scheme, EM-EP converges rapidly to a stationary point. To exploit discrete constellations, a decentralized EP-based algorithm is developed and further simplified using LSL approximations. To reduce the computational complexity of EP caused by constellation size, a hybrid Variational Bayes (VB) and EP method (VB-EP) is proposed. While VB-EP follows a similar design philosophy to EM-EP and variable-level EP (VL-EP), it introduces additional constraints that may degrade performance from the perspective of BFE optimization. In contrast, hierarchical EP avoids introducing new constraints and reduces computational complexity by exploiting the hierarchical structure of 4^M -QAM constellations. Motivated by emerging machine-type communications in 6G systems, grant-free (GF) CF networks are also investigated. A distributed iterative maximum likelihood algorithm is proposed for user activity detection. EP serves as the core framework for both linear and bilinear estimation in this thesis. A fundamental challenge in EP is the presence of non-integrable beliefs. In bilinear estimation, this issue arises when computing Gaussian-equivalent messages for Gaussian mixture model (GMM) messages. To address this, an EP-based algorithm is proposed to approximate the mean and variance of a random variable whose distribution is given by a product of GMMs. Two approaches are developed to preserve non-integrable

messages while preventing non-integrable beliefs. These approaches are further generalized into an EP framework capable of handling more complex factorization structures, such as those encountered in CF communication systems.

List of Figures

3.1	Performance of GLM-reVAMP algorithm with different variable length N .	35
5.1	Normalized l_2 norm of $\hat{\mathbf{x}}^t - \hat{\mathbf{x}}_{\text{MMSE}}$	58
5.2	Normalized l_1 norm of $\boldsymbol{\tau}_x^t - \boldsymbol{\tau}_{\text{MMSE}}$	58
6.1	NMSE [dB] versus SNR [dB]	70
7.1	Partial factor graph	73
7.2	NMSE vs SNR	80
8.1	Channel NMSE (dB) versus SNR (dB)	90
9.1	Factor Graph with Neighborhoods of both Factor (above) and Variable (below) Nodes. For simplicity, $\forall$ symbol is used to indicate that there are multiple nodes.	94
9.2	NMSE vs SNR	98
10.1	Comparison between PP-VB-EP and JACD-EP	110
11.1	Factor Graph for VDBP	116
11.2	Factor Graph for EP	120
11.3	NSE CDF of mean estimates.	130
11.4	NSE CDF of variance estimates.	130
12.1	CDF of NMSE	145

Chapter 1

Introduction

Cell-Free (CF) Massive MIMO (MaMIMO) is a promising technology for the next-generation communication networks. It is expected to combine the benefits of both MaMIMO and distributed antenna system (DAS) [1].

As suggested in [1], in CF-MaMIMO networks, a user equipment (UE) is jointly served by all the access points (APs) in a region. Each AP is equipped with either a single [1] or a very small number [2] of antennas. The APs cooperate in a phase-coherent manner via a backhaul network using time-division duplexing. Due to channel reciprocity, the downlink communication is only possible when the channel is estimated during the uplink. In this thesis, we focus on the uplink communication, particularly channel estimation. To limit backhaul transmission usage, the channel state information should not be shared among different APs [1].

A major challenge in CF networks is pilot contamination. Since all the users in a region are simultaneously served by all the APs in the region, the number of UEs may exceed the length of pilot sequences. As a result, the likelihood obtained from pilot observations is not sufficient to estimate the channel. Other information such as channel prior or data based estimation is needed to estimate the channel correctly. Based on these considerations, we divide the thesis into two parts. In the first part, we consider only prior information and the pilot observation. In this part, we generalize the pilot-based channel estimation into a sparse signal recovery problem, where the pilot sequences sent by all the UEs can be identified as the measurement matrix, and the channel the channel is treated as “unknown input signal” in the context of sparse signal recovery. Since the APs are distributed across the region, the channel between users and different APs can be considered independent. Therefore, if we consider single-antenna APs, we only need to investigate sparse vector signal recovery, instead of the whole channel matrix. In the case of multi-antenna APs, since every AP is only equipped with a small number of antenna, we can utilize vectorization to transform the problem into a vector signal recovery problem without significantly increasing the complexity. Therefore, in the first part, we investigate methods for sparse signal recovery.

In the second part, we estimate the channel using not only channel priors and pilot-based observations but also data-based observations. This leads to the so-called semi-blind estimation and can be formulated as a bilinear estimation (and detection) problem. By combining the pilot observation likelihood and the channel prior, we obtain a joint distribution of the received pilot signals and the channel. This allows us to transform the semi-blind channel estimation with independent (per-AP-per-UE) channel prior to a blind channel estimation problem with a hypothetical correlated channel prior (i.e.,

the channel of different users becomes correlated). To avoid excessive complications, we consider orthogonal pilot sequences, i.e., the inner product of the pilot sequences from two users is zero if they use different sequences, but they are also allowed to reuse the same sequence. It has been shown in [3] that using randomly assigned non-orthogonal pilot sequence outperforms orthogonal pilot assignment. Nevertheless, the focus of the second part is bilinear estimation, and the techniques developed in the first part can be readily extended to accommodate non-orthogonal pilots. With the growing interest of IoT and autonomous driving, grant-free (GF) communication is also required. It has been shown that MaMIMO systems are well suited to support GF access [4]. This thesis also includes a discussion of Grant Free CF-MaMIMO, in which not only the channel and data need to be estimated and detected, but also the user activities need to be detected.

1.1 Prior Works

1.1.1 Useful Bayesian Inference Tools

Both the linear and estimation problems can be generalized as high-dimensional Bayesian inference problems. Obtaining the exact optimal solution requires NP-hard computations to evaluate high-dimensional integrals and all possible discrete combinations. More practical approaches aim to obtain approximate solutions using iterative methods. In this thesis, we mainly employ message passing based methods [5–7]. These algorithms exploit the structure of the factored joint probability distribution to compute the marginal distributions. From a factored joint probability, a factor graph can be constructed. A factor graph consists of two types of nodes: factor nodes and variable nodes. These nodes can be obtained by listing all the factors and variables (only once, even if a variable appears in multiple factors). An edge is drawn between a factor node and a variable node if the variable is an argument of the factor.

If the factor graph contains no loops, then sum-product rules [5] can be applied to iteratively compute marginal distributions. This method can be shown recursively to yield exact marginals. The sum-product rules depend only on the local neighborhoods of each factor and variable node, and therefore can also be applied to a loopy factor graphs. This idea is called Loopy Belief Propagation (BP) and is studied in [8]. At each factor node, a belief is computed, which is normalized to 1 and proportional to the product of the factor and all incoming messages from neighboring variable nodes. These incoming messages are referred to as extrinsics to that factor. Similarly, at each variable node, a belief is obtained as the product of all incoming messages to that node from neighboring factor nodes. In BP, the message from factor to variable is computed in two steps:

- Marginalize the belief so that it depends only on the destination variable.
- Compute the message from the factor to the variable as the quotient of the marginalized belief and the message from that variable to the factor.

Since the belief at variable already depends on only a single variable, no marginalization is needed. Therefore, the message from a variable to a factor can be directly computed as the quotient of the variable-level belief and the incoming message from the factor. At convergence, these beliefs can be interpreted as approximate marginal distributions, and they are exact if the factorization is loop-free. Although BP decomposes a global inference problem into multiple local ones, multiple arbitrary messages at a linear combination

factor can still lead to intractable computations, as is the case in CF networks. As will be shown in the second part of this thesis, the messages from channel coefficients (variables) to the pilot observation likelihood (factor) take the form of Gaussian mixture models (GMMs). Even with orthogonal pilots, the marginalization step at the observation likelihood remains intractable.

Expectation Propagation (EP) [7,9] introduces a further approximation to relax the problem. The goal of EP is to restrict messages to a chosen family of functions. Compared to the BP rules mentioned before, EP only introduces an additional projection operation when computing the message from factor to variable. In EP, the message from factor to variable is summarized in the following three steps:

- Marginalize the belief such that it depends only on the destination variable.
- Project the resulting marginal belief onto desired family of distributions. In this thesis, the projection is always onto the Gaussian family.
- Compute the message from the factor to the variable as the quotient of the projected marginalized belief and the incoming message from that variable.

This additional simplification significantly expands the applicability of message-passing algorithms. However, a major limitation of EP is that the resulting messages may be non-integrable, even when all factors are integrable. This issue does not arise in BP, since the quotient operation effectively removes one component from a product. The presence of non-integrable messages is a major drawback of EP, as it may prevent the algorithm from proceeding. This remains an open problem [10], which we address in the second part of this thesis.

Constrained Bethe Free Energy (BFE) optimization provides a unified framework for message-passing algorithms such as BP and EP [6,11]. Given a factorization of the original joint distribution, BFE represents the approximate variational free energy under the Bethe approximation. The concepts of variational free energy and the Bethe approximation are briefly introduced below:

- Variational Free Energy is essentially the Kullback-Leibler divergence (KLD) between a trial distribution and the given factorized distribution, where the trial distribution appears as the first argument of the KLD.
- Bethe approximation: A loop-free factorized joint distribution can be expressed in a quotient form involving only marginal distributions. The numerator consists of multivariate marginal distributions, each involving the variables of a corresponding factor. The denominator consists of variable-level marginal distributions. The Bethe approximation defines a trial distribution that preserves this quotient structure and, consequently, the local dependencies. The trial factors in the numerator can be interpreted as factor-level beliefs, while those in the denominator correspond to variable-level beliefs in BP or EP.

If a loop-free factorization is assumed, enforcing that each trial factor in the denominator equals the marginalization of the corresponding numerator factor yields exact marginals. Furthermore, BP and EP can be unified within the constrained BFE optimization framework [6,11] by imposing different consistency constraints between the numerator and denominator of the Bethe approximation. Strict marginal consistency constraints correspond to BP, whereas relaxed moment consistency constraints correspond to EP.

1.1.2 Prior Works for Linear Estimation

We can formulate pilot-based channel estimation in CF networks as a sparse signal recovery problem. Approximate Message Passing (AMP) is a well-known computationally efficient algorithm for sparse signal recovery in linear models [12–14]. The authors show in [12] that AMP can be derived from BP with large-system simplifications based on Taylor expansions. A main feature of AMP is that, by using large-system-limit (LSL) approximations, AMP is able to represent messages as functions of beliefs. The same idea can be extended to generalized linear models [15, 16], leading to the Generalized AMP (GAMP) algorithm.

In [15], the authors show that GAMP can be derived based on EP and LSL approximations. Since both BP and EP are closely related to BFE, the connection between BFE and (G)AMP is studied in [17, 18]. In [17], the authors show that the (G)AMP algorithm has the same stationary points as the BFE under large-system-limit approximations. The convergence of message-passing algorithms such as BP, EP, and GAMP remains an open problem. The dynamics of AMP can only be computed numerically as a one-dimensional state evolution [12, 13]. Following the observation that GAMP and LSL-BFE share the same stationary points, Rangan proposed an analytically provable convergent method based on the optimization of the LSL-BFE using ADMM [18]. However, a disadvantage of ADMM-based methods is their cubic complexity due to matrix inversion operations. As shown in [13], the AMP algorithm is very sensitive to the measurement matrix, and the large-system limit does not hold for highly structured measurement matrices. Since (G)AMP can be interpreted as BP or EP with large-system-limit approximations, [19] proposes an EP-like algorithm, Vector AMP (VAMP), to handle signal recovery problems with structured measurement matrices.

1.1.3 Prior Works for Bilinear Estimation

The major challenge in bilinear estimation problems is that the product of two random variables is often non-trivial. For example, if we consider a Gaussian channel with Gaussian signaling, the resulting product random variable takes the form of Meijer G-functions. In more practical cases, the channel is Gaussian and the user signal follows a discrete alphabet, in which case the resulting product random variable is a Gaussian Mixture Model (GMM). Although linear combinations of GMMs remain GMMs, the number of Gaussian components in the resulting GMM grows exponentially with the number of factor GMMs. Therefore, iterative methods are necessary to make CF communications practical.

Variable-Level Expectation Propagation (VL-EP) [20] mitigates the intractable computations in bilinear estimation using iterative Expectation Maximization (EM). In VL-EP, interference is treated as Gaussian due to the Central Limit Theorem (CLT). Moreover, to remove the effect of interference from other users when estimating the channel of a given user, VL-EP uses channel posteriors obtained from previous iterations, similar to decision feedback. Subsequently, EM is applied, where user data are treated as latent variables. Therefore, VL-EP can be interpreted as an iterative method for finding an approximate maximum a posteriori estimate of the channel. A major limitation of VL-EP is that it cannot be deployed in a distributed manner, since the “expectation” step in the EM algorithm requires previously estimated channel coefficients from all APs.

In [21, 22], the authors propose Hybrid Vector Message Passing (HVMP). HVMP transforms the bilinear estimation problem into iterative linear estimation problems using a variational Bayes (VB)-like approach, which assumes that the channel posterior is inde-

pendent of the data posterior. This assumption can be incorporated into the constrained BFE optimization framework by introducing a constraint on the joint posterior of the channel and the data, such that it factorizes into the product of the channel posterior and the data posterior. As a result, the bilinear estimation problem reduces to iterative alternating linear estimation problems. The authors then apply AMP to perform the linear estimation. The core computation of HVMP is very similar to that of VL-EP, as both approaches convert bilinear estimation into linear estimation by using the expectation of the log joint distribution.

In [23], the authors propose a bilinear version of GAMP algorithm (BiGAMP) based on BP, CLT, and Taylor approximations under LSL. Similar to (G)AMP in linear systems, BiGAMP represents messages as functions of beliefs. However, due to second-order Taylor expansions, the messages in BiGAMP take the form of compositions of exponential and quadratic functions. Unlike Gaussian functions, these messages are not guaranteed to be integrable.

In practical communication scenarios, user data symbols are drawn from discrete constellations. To exploit this structure, [2] proposes an EP-based algorithm for joint channel estimation and data detection (referred to as bilinear-EP JCD). By combining EP with BP, messages related to data symbols are computed using BP, thereby preserving the discrete nature of the data symbols, while messages related to the channel are projected onto Gaussian distributions using EP.

1.2 Main Contributions

The thesis is mainly divided into two parts. In the first part, we focus on estimation techniques in generalized linear models. In the second part, we focus on bilinear estimation techniques.

1.2.1 Main Contributions in Linear Estimations

The VAMP algorithm [19] can be used for signal recovery in linear models. However, it only estimates average posterior variances. To address this limitation, we develop a revisited VAMP algorithm for generalized linear models (reGVAMP), which is not only applicable to generalized linear models but also computes component-wise posterior variances. The reGVAMP algorithm further provides a method for constructing equivalent Gaussian priors and likelihoods, thereby transforming a generalized linear model with arbitrary input priors and output observation likelihoods into an equivalent linear model with Gaussian priors and additive white Gaussian noise (AWGN), while preserving the posterior means and variances.

We then analyze the fixed points of reGVAMP by establishing the corresponding constrained BFE optimization problem. We provide an alternative derivation to obtain the large-system-limit (LSL) simplification of this constrained BFE formulation. Furthermore, we show that the extrinsic message to the data prior factor corresponds to the component-wise conditionally unbiased LMMSE (CWCU LMMSE) [24, 25] estimate of the equivalent linear model.

Based on the analysis of the resulting LSL-BFE, we propose a convergent AMBGAMP algorithm. To address the matrix inversion required by the least-squares steps in ADMM-GAMP [18] and AMBGAMP, we introduce a line-search-based gradient descent method.

Since the proposed method deviates from the ADMM framework, a sketch of the convergence proof is also provided.

1.2.2 Main Contributions in Bilinear Estimations

The VL-EP algorithm uses posterior estimates for interference cancellation. However, this approach does not support distributed implementation, since the APs must share their estimated channel coefficients with the CPU for the expectation step. To address this limitation, we develop an EP framework combined with EM to handle bilinear estimation. Due to the fact that the operation $\exp(\mathbb{E}[\log(\cdot)])$ transforms the bilinear likelihood into a linear Gaussian likelihood in the EM algorithm, the messages in the proposed EM-EP framework are always integrable. According to the philosophy of reGVAMP, the message from a variable to the observation likelihood can be interpreted as the approximate marginalization of a partial distribution involving all other factors. Therefore, in EM-EP, the partial distributions instead of the posteriors are used for interference cancellation. This EM-EP framework enables flexible switching between centralized and distributed implementations by adjusting the factorization scheme, provided that the channel priors and noise across different APs are independent. Since loops in the factorization scheme affect the convergence of the underlying message-passing algorithm, we further investigate EM-EP under a loop-free factorization scheme.

By utilizing the LSL properties described in [13], we investigate a simplified Bilinear EP algorithm based on LSL approximations. From the ideas of consensus propagation [26,27], the reliance on a central processing unit (CPU) can be eliminated, thereby enabling the transformation of a distributed system into a fully decentralized system.

According to [6, 11], message-passing algorithms are closely related to constrained BFE optimization. We investigate the BFE formulation for CF networks and propose a hybrid VB-EP algorithm by introducing a constraint that approximates the joint belief (posterior) of the channel and data for each user as independent beliefs. This approach is similar to the work in [21]. However, the prior work in [21] assumes that the joint belief of the entire channel and data matrix is fully factorized. In contrast, we aim to preserve the dependencies between the channel and data of different users by introducing an auxiliary variable for the user-wise bilinear combination. The resulting algorithm employs expectation and logarithmic operations to convert the bilinear estimation problem into iterative linear estimation problems, similar to VB and VL-EP. Compared to the EP-based algorithm described above and [2], where each AP must evaluate a GMM with a number of Gaussian components equal to the constellation size, VB-EP requires only the mean and variance of the user data beliefs. This enables the transformation of the bilinear problem into linear estimation problems involving only simple Gaussian distributions.

Although VB-EP reduces the computational complexity at the AP side, introducing additional constraints inevitably leads to performance degradation. We observe that when square-shaped 4^M -QAM constellations are used, they can be decomposed into M subconstellations, each corresponding to a 4-PSK constellation. We exploit this property and propose a hierarchical EP method for bilinear estimation. By exploiting the hierarchical structure and decomposing QAM into PSK components, we reduce the computational complexity, especially in multi-antenna AP scenarios. This is because the covariance matrices of the Gaussian components in the channel belief GMM depend on the squares of the symbol realizations. Since the square of a PSK symbol is constant, the matrix inversions of these covariance matrices only need to be computed once.

We further consider machine-type communication by investigating grant-free (GF) access in CF MaMIMO systems. The key difference between conventional CF networks and GF-CF lies in the introduction of an activity indicator matrix. To address this problem, we propose a framework that combines deterministic detection with Bayesian estimation, where the activity indicator matrix is treated as an unknown deterministic variable. Specifically, we develop a distributed iterative maximum likelihood method to detect user activity. Once the user activity is identified, the GF-CF system reduces to a standard CF communication problem. In this thesis, VB-EP is then applied for bilinear estimation and detection in the resulting CF system.

Although all factorization schemes considered in this thesis consist of integrable factors, non-integrable messages (intuitive Gaussian with “negative” variance) may still prevent EP-based algorithms from proceeding, since non-integrable beliefs cannot be normalized and their means and variances do not exist. This issue is common in EP, and no general solution exists when non-integrable beliefs arise. In most existing works [2, 10, 19, 28], non-integrable beliefs are avoided by constraining the messages to remain integrable. However, from the perspective of constrained BFE optimization, restricting non-integrable messages reduces the feasible set of the optimization problem. For example, in a one-dimensional linear model with a BPSK prior, it can be verified that MMSE estimates are achievable only when non-integrable messages are allowed. In CF-MaMIMO systems with Gaussian channel priors and discrete data symbols, the channel belief at the bilinear combination factor becomes a GMM, which does not guarantee integrable feedback messages. In this thesis, we first isolate this subproblem—estimating the mean and variance of a random variable whose distribution is given by a product of GMMs. Simulation results show that allowing non-integrable messages significantly improves performance. We then generalize this idea and propose modified EP rules that prevent non-integrable beliefs while preserving the flexibility of the message updates.

Chapter 2

Detailed Introduction to BFE and Message Passing Algorithms

2.1 Bethe Free Energy

Consider a factored joint PDF:

$$p(\boldsymbol{\theta}) \propto \prod_{\alpha} f_{\alpha}(\boldsymbol{\theta}_{\alpha}). \quad (2.1)$$

The Bethe approximation for (2.1) has the trial function $b(\boldsymbol{\theta})$ of the form of:

$$b(\boldsymbol{\theta}) = \frac{\prod_{\alpha} b_{f_{\alpha}}(\boldsymbol{\theta}_{\alpha})}{\prod_i b_{\theta_i}(\theta_i)^{L_i-1}}, \quad (2.2)$$

where L_i denote the number of factors that include θ_i as a parameter. All the functions in (2.2) must be proper distributions integrate to one. Furthermore, the numerators in (2.2) must be consistent with the denominators, i.e.,

$$\begin{aligned} \forall \alpha \text{ and } i \in N(\alpha) : b_{\theta_i}(\theta_i) &= \int b_{f_{\alpha}}(\boldsymbol{\theta}_{\alpha}) \prod_{j \in N(\alpha)/\{i\}} d\theta_j, \\ \forall i : \int b_{\theta_i}(\theta_i) d\theta_i &= 1, \end{aligned} \quad (2.3)$$

where we exploit the notations and denote $N(\alpha)$ as the set of indices of entries contained in $\boldsymbol{\theta}_{\alpha}$. The Bethe Free Energy is actually, the variational energy of the Bethe Approximation:

$$\text{BFE} = \text{KLD}[b(\boldsymbol{\theta}) \| p(\boldsymbol{\theta})] \quad (2.4)$$

In order to determine all the $b_{f_{\alpha}}$ and b_{θ_i} , we minimize the Bethe Free Energy,

$$\begin{aligned} \min_{\forall \alpha, i : b_{f_{\alpha}}, b_{\theta_i}} \quad & \text{BFE} \\ \text{s.t.} \quad & (2.3). \end{aligned}$$

2.1.1 Tree Structured Factorization Example

Observe a simple example, assume a PDF can be decomposed into two factors:

$$p(\boldsymbol{\theta}) = \frac{1}{Z} f_{\alpha}(\boldsymbol{\theta}_{\alpha}) f_{\beta}(\boldsymbol{\theta}_{\beta}), \quad (2.5)$$

where

$$\begin{aligned}\boldsymbol{\theta} &= [\theta_1 \ \theta_2 \ \theta_3]^T \\ \boldsymbol{\theta}_\alpha &= [\theta_1 \ \theta_2]^T \\ \boldsymbol{\theta}_\beta &= [\theta_2 \ \theta_3]^T; \\ Z &= \int f_\alpha(\boldsymbol{\theta}_\alpha) f_\beta(\boldsymbol{\theta}_\beta) d\boldsymbol{\theta}.\end{aligned}\tag{2.6}$$

To verify the conditioned independence, we look at the relation between $p(\theta_1, \theta_3|\theta_2)$ and the product $p(\theta_1|\theta_2)p(\theta_3|\theta_2)$. According to Bayes rule, the conditional joint PDF is

$$p(\theta_1, \theta_3|\theta_2) = \frac{1}{Zp(\theta_2)} f_\alpha(\theta_1, \theta_2) f_\beta(\theta_2, \theta_3),\tag{2.7}$$

with the relation

$$\begin{aligned}Zp(\theta_2) &= \int f_\alpha(\boldsymbol{\theta}_\alpha) f_\beta(\boldsymbol{\theta}_\beta) d\theta_1 d\theta_3 \\ &= \int f_\alpha(\theta_1, \theta_2) d\theta_1 \cdot \int f_\beta(\theta_2, \theta_3) d\theta_3\end{aligned}\tag{2.8}$$

For simplicity, we define

$$\begin{aligned}Z_{\alpha|2} &= \int f_\alpha(\theta_1, \theta_2) d\theta_1 \\ Z_{\beta|2} &= \int f_\beta(\theta_2, \theta_3) d\theta_3.\end{aligned}\tag{2.9}$$

Therefore, the conditional joint PDF can be rewritten as

$$p(\theta_1, \theta_3|\theta_2) = \frac{1}{Z_{\alpha|2}Z_{\beta|2}} f_\alpha(\theta_1, \theta_2) f_\beta(\theta_2, \theta_3)\tag{2.10}$$

On the other hand, the conditional marginal PDFs can be computed as

$$\begin{aligned}p(\theta_1|\theta_2) &= \frac{Z_{\beta|2}}{Zp(\theta_2)} f_\alpha(\theta_1, \theta_2) = \frac{1}{Z_{\alpha|2}} f_\alpha(\theta_1, \theta_2) \\ p(\theta_3|\theta_2) &= \frac{1}{Z_{\beta|2}} f_\beta(\theta_2, \theta_3).\end{aligned}\tag{2.11}$$

As we can see, since $p(\theta_1, \theta_3|\theta_2) = p(\theta_1|\theta_2)p(\theta_3|\theta_2)$, the variables θ_1 and θ_3 are conditional independent given θ_2 . Therefore, regardless the form of function f_α and f_β , we have

$$p(\boldsymbol{\theta}) = p(\theta_1, \theta_3|\theta_2)p(\theta_2) = \frac{p(\theta_1, \theta_2)p(\theta_2, \theta_3)}{p(\theta_2)}\tag{2.12}$$

2.1.2 Analysis of Tree Structured Factorization

Lemma 1. *For all PDF $p(\boldsymbol{\theta})$ with $\forall i, \theta_i$ treated as atomic variables. If $p(\boldsymbol{\theta})$ admits a loop-free (tree/forest structure) factorization:*

$$p(\boldsymbol{\theta}) \propto \prod_{\alpha=1}^A f_\alpha(\boldsymbol{\theta}_\alpha),\tag{2.13}$$

then we have

$$p(\boldsymbol{\theta}) = \frac{\prod_{\alpha} p(\boldsymbol{\theta}_{\alpha})}{\prod_i p(\theta_i)^{L_i-1}}, \quad (2.14)$$

where L_i denotes the number of factors f_{α} that contains θ_i as a parameter.

Proof. We use mathematical induction on the number of factors.

If there is only one factor, the lemma holds automatically. Now assume the induction hypothesis. Assume that the lemma holds for all non-cyclic factored PDFs with equal or less than A factors.

We need to prove that the lemma also hold for any PDF with $A + 1$ non-cyclic structured factors:

$$p(\boldsymbol{\theta}) \propto \prod_{\alpha=1}^{A+1} f_{\alpha}(\boldsymbol{\theta}_{\alpha}). \quad (2.15)$$

Heuristically, there are two ways of decreasing the number of factors, either by conditioning or by marginalization and in this proof, we use the technique of marginalization. Since all the $A + 1$ factors compose a non-cyclic structure, there must exist a factor (leaf factor) that has at most 1 common atomic variable with the superfactor of the product of all the other factors. Without loss of generality, we assume $f_{A+1}(\boldsymbol{\theta}_{A+1})$ has at most one common atomic variable with $\prod_{\alpha=1}^A f_{\alpha}(\boldsymbol{\theta}_{\alpha})$ (we can always change the index if not). Therefore, we first investigate the PDF:

$$p(\boldsymbol{\theta}) \propto f_{A+1}(\boldsymbol{\theta}_{A+1}) \prod_{\alpha=1}^A f_{\alpha}(\boldsymbol{\theta}_{\alpha}). \quad (2.16)$$

If $f_{A+1}(\boldsymbol{\theta}_{A+1})$ is completely disjoint to $\prod_{\alpha=1}^A f_{\alpha}(\boldsymbol{\theta}_{\alpha})$ and zero common atomic variable with it, then by marginalization, we have

$$\begin{aligned} p(\boldsymbol{\theta}^-) &\propto \prod_{\alpha=1}^A f_{\alpha}(\boldsymbol{\theta}_{\alpha}) \\ \Rightarrow p(\boldsymbol{\theta}^-) &= \frac{\prod_{\alpha=1}^A p(\boldsymbol{\theta}_{\alpha})}{\prod_i p(\theta_i)^{L_i^- - 1}} \end{aligned} \quad (2.17)$$

where $\boldsymbol{\theta}^-$ is composed of the atomic variables in $\forall \alpha \in \{1, \dots, A\} : \boldsymbol{\theta}_{\alpha}$, L_i^- denotes the number of factors in $\{f_{\alpha}(\boldsymbol{\theta}_{\alpha}) | \alpha \in \{1, \dots, A\}\}$ that contain θ_i as parameter. The second line of (2.17) correspond to the induction hypothesis. On the other hand, we have

$$p(\boldsymbol{\theta}_{A+1}) \propto f_{A+1}(\boldsymbol{\theta}_{A+1}). \quad (2.18)$$

From (2.18) and the first line of (2.17), we find that (2.16) has the form of

$$\begin{aligned} p(\boldsymbol{\theta}) &= p(\boldsymbol{\theta}^-) p(\boldsymbol{\theta}_{A+1}) \\ &= \frac{\prod_{\alpha=1}^{A+1} p(\boldsymbol{\theta}_{\alpha})}{\prod_i p(\theta_i)^{L_i-1}} \end{aligned} \quad (2.19)$$

where L_i denotes the number of factors in (2.15) that contain θ_i as a parameter. Since there is no common atomic variable between $f_{A+1}(\boldsymbol{\theta}_{A+1})$ and $\prod_{\alpha=1}^A f_{\alpha}(\boldsymbol{\theta}_{\alpha})$, we notice the relation $L_i = L_i^-$.

Next we consider the case where $f_{A+1}(\boldsymbol{\theta}_{A+1})$ shares one common atomic variable with $\prod_{\alpha=1}^A f_{\alpha}(\boldsymbol{\theta}_{\alpha})$. We assume they share the c -th atomic variable as the common atomic variable, which we denote as θ_c . Similar to the example in Section 2.1.1, we can show by integration rules:

$$\begin{aligned} p(\bar{\boldsymbol{\theta}}_{A+1}|\theta_c) &= \frac{1}{Z_{\bar{\boldsymbol{\theta}}_{A+1}|\theta_c}} f_{A+1}(\boldsymbol{\theta}_{A+1}) \\ p(\bar{\boldsymbol{\theta}}^-|\theta_c) &= \frac{1}{Z_{\bar{\boldsymbol{\theta}}^-|\theta_c}} \prod_{\alpha=1}^A f_{\alpha}(\boldsymbol{\theta}_{\alpha}), \end{aligned} \quad (2.20)$$

where $\bar{\boldsymbol{\theta}}_{A+1}$ denotes $\boldsymbol{\theta}_{A+1}$ with θ_c removed, and $\bar{\boldsymbol{\theta}}^-$ denotes $\boldsymbol{\theta}^-$ with θ_c removed. Furthermore, $Z_{\bar{\boldsymbol{\theta}}_{A+1}|\theta_c}$ and $Z_{\bar{\boldsymbol{\theta}}^-|\theta_c}$ are normalization constants to ensure $p(\bar{\boldsymbol{\theta}}_{A+1}|\theta_c)$ and $p(\bar{\boldsymbol{\theta}}^-|\theta_c)$ integrate to 1. On the other hand, from (2.15), the conditional joint PDF given θ_c can also be verified to satisfy the following relation:

$$\begin{aligned} p(\bar{\boldsymbol{\theta}}|\theta_c) &= \frac{1}{Z_{\bar{\boldsymbol{\theta}}_{A+1}|\theta_c}} \frac{1}{Z_{\bar{\boldsymbol{\theta}}^-|\theta_c}} \prod_{\alpha=1}^{A+1} f_{\alpha}(\boldsymbol{\theta}_{\alpha}) \\ &= p(\bar{\boldsymbol{\theta}}_{A+1}|\theta_c) p(\bar{\boldsymbol{\theta}}^-|\theta_c), \end{aligned} \quad (2.21)$$

where $\bar{\boldsymbol{\theta}}$ denote $\boldsymbol{\theta}$ with θ_c removed. Due to the equality in the second line of (2.21), we know $\bar{\boldsymbol{\theta}}_{A+1}$ and $\bar{\boldsymbol{\theta}}^-$ are conditionally independent given θ_c . From (2.21) and Bayes rule, we have

$$\begin{aligned} p(\bar{\boldsymbol{\theta}}|\theta_c) p(\theta_c) &= p(\bar{\boldsymbol{\theta}}_{A+1}|\theta_c) p(\bar{\boldsymbol{\theta}}^-|\theta_c) \frac{p(\theta_c)}{p(\theta_c)} p(\theta_c) \\ \Rightarrow p(\boldsymbol{\theta}) &= \frac{p(\boldsymbol{\theta}_{A+1}) p(\boldsymbol{\theta}^-)}{p(\theta_c)}. \end{aligned} \quad (2.22)$$

We look at the marginal PDF $p(\boldsymbol{\theta}^-)$. From (2.16), we have

$$p(\boldsymbol{\theta}^-) \propto \prod_{\alpha=1}^A f_{\alpha}(\boldsymbol{\theta}_{\alpha}) \int f_{A+1}(\boldsymbol{\theta}_{A+1}) d\bar{\boldsymbol{\theta}}_{A+1}. \quad (2.23)$$

Since θ_c is the common atomic variable between $f_{A+1}(\boldsymbol{\theta}_{A+1})$ and $\prod_{\alpha=1}^A f_{\alpha}(\boldsymbol{\theta}_{\alpha})$, we know $\exists \beta \in \{1, \dots, A\}$, such that θ_c is a parameter of $f_{\beta}(\boldsymbol{\theta}_{\beta})$. We can define a new auxiliary factor $f'_{\beta}(\boldsymbol{\theta}_{\beta})$ such that

$$f'_{\beta}(\boldsymbol{\theta}_{\beta}) = f_{\beta}(\boldsymbol{\theta}_{\beta}) \int f_{A+1}(\boldsymbol{\theta}_{A+1}) d\bar{\boldsymbol{\theta}}_{A+1}. \quad (2.24)$$

Therefore, the marginal PDF $p(\boldsymbol{\theta}^-)$ can be written as

$$p(\boldsymbol{\theta}^-) \propto f'_{\beta}(\boldsymbol{\theta}_{\beta}) \prod_{\alpha=1}^{\beta-1} f_{\alpha}(\boldsymbol{\theta}_{\alpha}) \prod_{\alpha=\beta+1}^A f_{\alpha}(\boldsymbol{\theta}_{\alpha}). \quad (2.25)$$

Since there are only A factors, according to the induction hypothesis, the marginal PDF $p(\boldsymbol{\theta}^-)$ can be written as

$$p(\boldsymbol{\theta}^-) = \frac{\prod_{\alpha=1}^A p(\boldsymbol{\theta}_{\alpha})}{p(\theta_c)^{L_c-2} \prod_{i \neq c} p(\theta_i)^{L_i-1}}, \quad (2.26)$$

where L_i and L_c denote the number of factors in (2.15) that use θ_i and θ_c as parameter, respectively. The reason we have $L_c - 2$ instead of $L_c - 1$ is due to the fact that there are only $L_c - 1$ factors in (2.25) that contain θ_c .

Finally, we substitute (2.26) into (2.22) and obtain

$$p(\boldsymbol{\theta}) = \frac{\prod_{\alpha=1}^{A+1} p(\boldsymbol{\theta}_\alpha)}{\prod_i p(\theta_i)^{L_i-1}}. \quad (2.27)$$

□

Lemma 2. *If $b(\boldsymbol{\theta})$ is defined as*

$$b(\boldsymbol{\theta}) = \frac{\prod_{\alpha=1}^A b_{f_\alpha}(\boldsymbol{\theta}_\alpha)}{\prod_i b_{\theta_i}(\theta_i)^{L_i-1}}, \quad (2.28)$$

where

- L_i denote the number of factors f_α that contains θ_i as parameter;
- $b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$ compose a non-cyclic factor graph
- b_{f_α} and b_{θ_i} are non-negative functions:

$$\begin{aligned} \forall \alpha, i \in N(\alpha) : b_{\theta_i}(\theta_i) &= \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \prod_{j \in N(\alpha)/\{i\}} d\theta_j, \\ \forall i : \int b_{\theta_i}(\theta_i) d\theta_i &= 1, \\ \forall \alpha : \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) d\boldsymbol{\theta}_\alpha &= 1, \end{aligned} \quad (2.29)$$

then we have

$$\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) = \int b(\boldsymbol{\theta}) \prod_{i \notin N(\alpha)} d\theta_i. \quad (2.30)$$

Proof. We use mathematical induction on the number of factors in the numerator. If there is only one factor, by (2.28) we have $b(\boldsymbol{\theta}) = b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$, where $\boldsymbol{\theta} = \boldsymbol{\theta}_\alpha$.

As induction hypothesis, we assume the lemma hold for A factors in the numerator, we need to prove the lemma also hold for $b(\boldsymbol{\theta})$ with $A + 1$ factors.

Since $b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$ compose a non-cyclic factor graph (tree or forest), there exist a factor $b_\beta(\boldsymbol{\theta}_\beta)$ such that $b_\beta(\boldsymbol{\theta}_\beta)$ share at most one common atomic variable θ_c with $\prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$.

If there are no common atomic variables, we rewrite (2.28):

$$b(\boldsymbol{\theta}) = \frac{b_{f_\beta}(\boldsymbol{\theta}_\beta) \prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)}{\prod_i b_{\theta_i}(\theta_i)^{L_i-1}}. \quad (2.31)$$

We integrate out $\boldsymbol{\theta}_\beta$, and obtain

$$b^-(\boldsymbol{\theta}^-) = \int b(\boldsymbol{\theta}) d\boldsymbol{\theta}_\beta = \frac{\prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)}{\prod_i b_{\theta_i}(\theta_i)^{L_i-1}}, \quad (2.32)$$

where $\boldsymbol{\theta}^-$ denote $\boldsymbol{\theta}$ with the ones in $\boldsymbol{\theta}_\beta$ removed. In (2.32), there are only A factors in the numerator. Due to the induction hypothesis, we have

$$\forall \alpha \in \{1, \dots, \beta - 1, \beta + 1, \dots, A + 1\} : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) = \int b(\boldsymbol{\theta}) \prod_{i \notin N(\alpha)} d\theta_i. \quad (2.33)$$

Furthermore, due to the hypothesis, from (2.32)-(2.33), we have

$$\int b^-(\boldsymbol{\theta}^-) d\boldsymbol{\theta}^- = 1. \quad (2.34)$$

From (2.31) and (2.32),

$$b(\boldsymbol{\theta}) = b_{f_\beta}(\boldsymbol{\theta}_\beta) b^-(\boldsymbol{\theta}^-). \quad (2.35)$$

Integrate out all the variables in $\boldsymbol{\theta}^-$, we have

$$\int b(\boldsymbol{\theta}) \prod_{i \notin N(\beta)} d\theta_i = b_{f_\beta}(\boldsymbol{\theta}_\beta). \quad (2.36)$$

Now we consider the case where there is one common atomic variable θ_c between $b_{f_\beta}(\boldsymbol{\theta}_\beta)$ and $\prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$. We integrate the variables that only occur once in $b_{f_\beta}(\boldsymbol{\theta}_\beta)$:

$$\begin{aligned} b^-(\boldsymbol{\theta}^-) &= \int b(\boldsymbol{\theta}) \prod_{i \in N(\beta)/\{c\}} d\theta_i \\ &= \frac{b_{\theta_c}(\theta_c) \prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)}{b_{\theta_c}(\theta_c)^{L_c-1} \prod_{i \neq c} b_{\theta_i}(\theta_i)^{L_i-1}} \\ &= \frac{\prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)}{b_{\theta_c}(\theta_c)^{L_c-2} \prod_{i \neq c} b_{\theta_i}(\theta_i)^{L_i-1}}. \end{aligned} \quad (2.37)$$

Since only θ_c occurs $L_c - 1$ times in $\forall \alpha \in \{1, \dots, \beta - 1, \beta + 1, \dots, A + 1\} : f_\alpha(\boldsymbol{\theta}_\alpha)$, all the other atomic variables $\forall i : \theta_i$ occur L_i times, we can use the induction hypothesis on (2.37) and obtain:

$$\forall \alpha \in \{1, \dots, \beta - 1, \beta + 1, \dots, A + 1\} : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) = \int b(\boldsymbol{\theta}) \prod_{i \notin N(\alpha)} d\theta_i. \quad (2.38)$$

On the other hand, from (2.37) and the lemma assumption,

$$\int b^-(\boldsymbol{\theta}^-) \prod_{i \notin N(\beta)} d\theta_i = b_{\theta_c}(\theta_c). \quad (2.39)$$

From (2.28) and (2.37), we can write

$$b(\boldsymbol{\theta}) = b_{f_\beta}(\boldsymbol{\theta}_\beta) \frac{b^-(\boldsymbol{\theta}^-)}{b_{\theta_c}(\theta_c)}. \quad (2.40)$$

We integrate out the variables that does not show up in $b_{f_\beta}(\boldsymbol{\theta}_\beta)$, and based on (2.39):

$$\int b(\boldsymbol{\theta}) \prod_{i \notin N(\beta)} d\theta_i = b_{f_\beta}(\boldsymbol{\theta}_\beta) \quad (2.41)$$

□

2.1.3 Bethe Free Energy in Non-Cyclic Graph

In the case of non-cyclic factorization, we substitute (2.3) into (2.4) and by Lemma 2

$$\text{BFE} = \sum_{\alpha} D[b_{f_{\alpha}}(\boldsymbol{\theta}_{\alpha}) \| f_{\alpha}(\boldsymbol{\theta}_{\alpha})] + \sum_i (L_i - 1) H[b_{\theta_i}(\theta_i)]. \quad (2.42)$$

If the constraints (2.3) are satisfied, the minimal of the BFE in (2.42) is achieved when $b(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$. Furthermore, according to Lemma 2, if $b(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ and the constraints (2.3) are satisfied, all the marginal beliefs $\forall \alpha \forall i : b_{f_{\alpha}}(\boldsymbol{\theta}_{\alpha}), b_{\theta_i}(\theta_i)$ are fixed.

2.1.4 Bethe Free Energy Extended to Graph with Cycles

If the factored PDF implies a factor graph with cycles, the above two lemmas no longer hold. However, we still approximate the BFE to the form of (2.42).

Furthermore, the strong marginal consistency constraints may lead to intractable computation in an iterative algorithm. To further simplify the problem, the consistency constraints can be relaxed to moment constraints. Therefore, instead of (2.3), we can use the following relaxed moment constraints

$$\begin{aligned} \forall \alpha : \int b_{f_{\alpha}}(\boldsymbol{\theta}_{\alpha}) d\boldsymbol{\theta}_{\alpha} &= 1 \\ \forall i : \int b_{\theta_i}(\theta_i) d\theta_i &= 1 \\ \forall \alpha \text{ and } i \in N(\alpha) : \mathbb{E}_{b_{\theta_i}}[\phi(\theta_i)] &= \mathbb{E}_{b_{f_{\alpha}}}[\phi(\theta_i)]. \end{aligned} \quad (2.43)$$

2.2 Relation Between Bethe Free Energy and Message Passing Algorithms

The constraint Bethe Free Energy minimization lead to an optimization problem. We take strong constraints as an example and use Lagrangian multiplier to gain an insight of the optimization problem. The Lagrangian function can be written as

$$\begin{aligned} \mathcal{L} = \sum_{\alpha} D[b_{f_{\alpha}}(\boldsymbol{\theta}_{\alpha}) \| f_{\alpha}(\boldsymbol{\theta}_{\alpha})] + \sum_i (L_i - 1) H[b_{\theta_i}(\theta_i)] + \\ + \sum_{\alpha} \sum_{i \in N(\alpha)} \int \lambda_{\theta_i; f_{\alpha}}(\theta_i) [b_{\theta_i}(\theta_i) - b_{f_{\alpha}}(\theta_i)] d\theta_i, \end{aligned} \quad (2.44)$$

where

$$b_{f_{\alpha}}(\theta_i) = \int b_{f_{\alpha}}(\boldsymbol{\theta}_{\alpha}) \prod_{j \in N(\alpha)/\{i\}} d\theta_j. \quad (2.45)$$

Setting the variational derivative to zero:

$$\begin{aligned} \forall \alpha : \frac{\partial \mathcal{L}}{\partial b_{f_{\alpha}}(\boldsymbol{\theta}_{\alpha})} &= \ln b_{f_{\alpha}}(\boldsymbol{\theta}_{\alpha}) - \ln f_{\alpha}(\boldsymbol{\theta}_{\alpha}) - \sum_{i \in N(\alpha)} \lambda_{\theta_i; f_{\alpha}}(\theta_i) + C := 0 \\ \forall i : \frac{\partial \mathcal{L}}{\partial b_{\theta_i}(\theta_i)} &= (1 - L_i) \ln b_{\theta_i}(\theta_i) + \sum_{\alpha \in N(i)} \lambda_{\theta_i; f_{\alpha}}(\theta_i) + C := 0, \end{aligned} \quad (2.46)$$

where C denote a normalization term for the beliefs due to the normalization constraints in (2.3):

$$\begin{aligned}\forall \alpha : \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) d\boldsymbol{\theta}_\alpha &= 1; \\ \forall i : \int b_{\theta_i}(\theta_i) d\theta_i &= 1.\end{aligned}\tag{2.47}$$

Rewrite (2.46) and we have

$$\begin{aligned}\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) &\propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(\alpha)} \mu_{\theta_i; f_\alpha}(\theta_i); \\ \forall i : b_{\theta_i}(\theta_i) &\propto \left[\prod_{\alpha \in N(i)} \mu_{\theta_i; f_\alpha}(\theta_i) \right]^{\frac{1}{L_1 - 1}},\end{aligned}\tag{2.48}$$

where we define

$$\mu_{\theta_i; f_\alpha}(\theta_i) \triangleq \exp(\lambda_{\theta_i; f_\alpha}(\theta_i)).\tag{2.49}$$

Lemma 3. *Let $\mathcal{A}$ denote any countable set of indices with cardinality $|\mathcal{A}| > 1$. If $\forall \alpha \in \mathcal{A}$, $\mu_{\theta; f_\alpha}(\theta)$ are defined as positive functions, then there exist a unique (exactly one) set of positive functions denoted by $\forall \alpha \in \mathcal{A}$: $\mu_{f_\alpha; \theta}(\theta)$, such that*

$$\forall \alpha \in \mathcal{A} : \mu_{\theta; f_\alpha}(\theta) = \prod_{\beta \in \mathcal{A}/\{\alpha\}} \mu_{f_\beta; \theta}(\theta).\tag{2.50}$$

Proof. We exploit the monotonic property of log operation. Due to the positive function assumption, at every point θ , we compute the log of both sides of (2.50),

$$\forall \alpha \in \mathcal{A} : \lambda_{\theta; f_\alpha}(\theta) = \sum_{\beta \in \mathcal{A}/\{\alpha\}} \lambda_{f_\beta; \theta}(\theta),\tag{2.51}$$

where

$$\begin{aligned}\forall \alpha \in \mathcal{A} : \lambda_{\theta; f_\alpha}(\theta) &\triangleq \log \mu_{\theta; f_\alpha}(\theta); \\ \forall \alpha \in \mathcal{A} : \lambda_{f_\alpha; \theta}(\theta) &\triangleq \log \mu_{f_\alpha; \theta}(\theta).\end{aligned}\tag{2.52}$$

Denote the elements of $\mathcal{A}$ as

$$\mathcal{A} = \{\alpha_1, \alpha_2, \dots, \alpha_{|\mathcal{A}|}\},\tag{2.53}$$

where $|\mathcal{A}|$ denote the cardinality of set $\mathcal{A}$. We can define the vectors

$$\begin{aligned}\boldsymbol{\lambda}_{\theta; f}(\theta) &\triangleq \begin{bmatrix} \lambda_{\theta; f_{\alpha_1}}(\theta) & \dots & \lambda_{\theta; f_{\alpha_{|\mathcal{A}|}}}(\theta) \end{bmatrix}^T, \\ \boldsymbol{\lambda}_{f; \theta}(\theta) &\triangleq \begin{bmatrix} \lambda_{f_{\alpha_1}; \theta}(\theta) & \dots & \lambda_{f_{\alpha_{|\mathcal{A}|}}; \theta}(\theta) \end{bmatrix}^T.\end{aligned}\tag{2.54}$$

With the above definition, (2.51) can be rewritten to

$$\boldsymbol{\lambda}_{\theta; f}(\theta) = \mathbf{B}_{|\mathcal{A}|} \boldsymbol{\lambda}_{f; \theta}(\theta),\tag{2.55}$$

where

$$\mathbf{B}_{|\mathcal{A}|} = \mathbf{1}_{|\mathcal{A}|} \mathbf{1}_{|\mathcal{A}|}^T - \mathbf{I}_{|\mathcal{A}|}.\tag{2.56}$$

Sylvester's determinant theorem, the absolute value of the determinant of $\mathbf{B}_{|\mathcal{A}|}$ can be computed as $|\det(\mathbf{B}_{|\mathcal{A}|})| = |\mathcal{A}| - 1$. Due to the lemma assumption, $|\mathcal{A}| > 1$, we know $\mathbf{B}_{|\mathcal{A}|}$ has full rank.

Therefore, from (2.55), the set of equations can be uniquely solved by

$$\boldsymbol{\lambda}_{f;\theta}(\theta) = \mathbf{B}_{|\mathcal{A}|}^{-1} \boldsymbol{\lambda}_{\theta;f}(\theta) = \left(\frac{\mathbf{1}_{|\mathcal{A}|} \mathbf{1}_{|\mathcal{A}|}^T}{N-1} - \mathbf{I}_{|\mathcal{A}|} \right) \boldsymbol{\lambda}_{\theta;f}(\theta). \quad (2.57)$$

□

Based on Lemma 3, we can uniquely define a set of positive functions $\forall i, \forall \alpha \in N(\theta_i) : \mu_{f_\alpha;\theta_i}(\theta_i)$, such that

$$\forall i, \forall \alpha \in N(\theta_i) : \mu_{\theta_i;f_\alpha}(\theta_i) = \prod_{\beta \in N(\theta_i)/\{\alpha\}} \mu_{f_\beta;\theta_i}(\theta_i). \quad (2.58)$$

Substitute (2.58) into the second line of (2.48), and (2.48) becomes

$$\begin{aligned} \forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) &\propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(\alpha)} \mu_{\theta_i;f_\alpha}(\theta_i); \\ \forall i : b_{\theta_i}(\theta_i) &\propto \prod_{\alpha \in N(\theta_i)} \mu_{f_\alpha;\theta_i}(\theta_i). \end{aligned} \quad (2.59)$$

Besides setting the derivative to zero, which results to (2.59), we must also include the consistency constraints (2.3). Therefore, by combining (2.3), (2.58) and (2.59), we obtain a system of equations (2.60) that corresponds to the stable points of the constrained BFE optimization problem:

$$\begin{aligned} \forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) &\propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(\alpha)} \mu_{\theta_i;f_\alpha}(\theta_i); \\ \forall i : b_{\theta_i}(\theta_i) &\propto \prod_{\alpha \in N(\theta_i)} \mu_{f_\alpha;\theta_i}(\theta_i); \\ \forall i, \forall \alpha \in N(\theta_i) : \mu_{\theta_i;f_\alpha}(\theta_i) &= \prod_{\beta \in N(\theta_i)/\{\alpha\}} \mu_{f_\beta;\theta_i}(\theta_i); \\ \forall \alpha \text{ and } i \in N(f_\alpha) : b_{\theta_i}(\theta_i) &= \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \prod_{j \in N(f_\alpha)/\{i\}} d\theta_j \end{aligned} \quad (2.60)$$

Actually, (2.60) is a system of equation parameterized by functions $\forall i, \forall \alpha \in N(\theta_i) : \mu_{f_\alpha;\theta_i}(\theta_i)$. Namely, every quantity appeared in (2.60) are uniquely determined, if $\forall i, \forall \alpha \in N(\theta_i) : \mu_{f_\alpha;\theta_i}(\theta_i)$ are given. The normalized non-negative function $b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$ and $b_{\theta_i}(\theta_i)$ are called belief at factor f_α and variable θ_i respectively. Moreover, the positive functions $\mu_{\theta_i;f_\alpha}(\theta_i)$ and $\mu_{f_\alpha;\theta_i}$ are called “messages from θ_i to f_α ” and “messages from f_α to θ_i ” respectively.

2.2.1 Relation to Belief Propagation (Strict Marginal Consistency Constraints)

On the other hand, belief propagation is an iterative message passing based algorithm by iterating over the following steps:

- Compute the belief at factor node:

$$\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(\alpha)} \mu_{\theta_i; f_\alpha}(\theta_i) \quad (2.61)$$

- Marginalize the belief at factor node:

$$\forall \alpha, \forall i \in N(f_\alpha) : b_{f_\alpha}(\theta_i) = \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \prod_{j \in N(f_\alpha)/\{i\}} d\theta_j \quad (2.62)$$

- Update the message from factors to variables:

$$\mu_{f_\alpha; \theta_i}(\theta_i) = \frac{b_{f_\alpha}(\theta_i)}{\mu_{\theta_i; f_\alpha}(\theta_i)} = \int f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{j \in N(f_\alpha)/\{i\}} \mu_{\theta_j; f_\alpha}(\theta_j) d\theta_j \quad (2.63)$$

- Compute the belief at variable node:

$$\forall i : b_{\theta_i}(\theta_i) \propto \prod_{\alpha \in N(\theta_i)} \mu_{f_\alpha; \theta_i}(\theta_i) \quad (2.64)$$

- Update the message from variable to factor:

$$\forall i, \forall \alpha \in N(\theta_i) : \mu_{\theta_i; f_\alpha}(\theta_i) = \frac{b_{\theta_i}(\theta_i)}{\mu_{f_\alpha; \theta_i}(\theta_i)}. \quad (2.65)$$

At convergence, the equations from (2.61)-(2.65) hold simultaneously, which compose a system of equations. Actually, one can show that the system of equations composed of (2.61)-(2.65) is equivalent to the system of equations (2.60), i.e.,

$$(2.61)-(2.65) \Leftrightarrow (2.60). \quad (2.66)$$

This relation indicates that Belief Propagation is just an iterative way of solving the system of equations in (2.60).

2.2.2 Relation to Expectation Propagation (Relaxed Moment Consistency Constraints)

Similar derivation can also be applied to the BFE minimization with relaxed moment constraints as indicated in (2.43). However, if relaxed consistency is used, we substitute the variational Lagrangian multiplier $\lambda_{\theta_i; f_\alpha}(\theta_i)$ with vector Lagrangian multiplier $\boldsymbol{\lambda}_{\theta_i; f_\alpha}$ (the same dimension as $\boldsymbol{\phi}(\theta_i)$). Therefore, finding the critical points of the Lagrangian function of BFE minimization with relaxed moment consistency constraints (2.43)

$$\begin{aligned} \forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) &\propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(f_\alpha)} \mu_{\theta_i; f_\alpha}(\theta_i); \\ \forall i : b_{\theta_i}(\theta_i) &\propto \prod_{\alpha \in N(\theta_i)} \mu_{f_\alpha; \theta_i}(\theta_i); \\ \forall i, \forall \alpha \in N(\theta_i) : \mu_{\theta_i; f_\alpha}(\theta_i) &= \prod_{\beta \in N(\theta_i)/\{\alpha\}} \mu_{f_\beta; \theta_i}(\theta_i); \\ \forall \alpha \text{ and } i \in N(f_\alpha) : \mathbb{E}_{b_{\theta_i}}[\boldsymbol{\phi}(\theta_i)] &= \mathbb{E}_{b_{f_\alpha}}[\boldsymbol{\phi}(\theta_i)], \end{aligned} \quad (2.67)$$

where all the variable-to-factor messages belongs to family $\mathcal{F}$, i.e.,

$$\forall \alpha, \forall i \in N(f_\alpha) \mu_{\theta_i; f_\alpha}(\theta_i) \in \mathcal{F}, \quad (2.68)$$

where

$$\mathcal{F} = \{\exp(\boldsymbol{\gamma}^T \boldsymbol{\phi}(\theta)) : \mathbb{C}^{\dim(\theta)} \rightarrow \mathbb{R} | \boldsymbol{\gamma} \in \mathbb{C}^{\dim(\phi(\theta))}\}. \quad (2.69)$$

On the other hand, the Expectation Propagation rules are

- Compute the belief at factor node:

$$\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(\alpha)} \mu_{\theta_i; f_\alpha}(\theta_i) \quad (2.70)$$

- Marginalize the belief at factor node:

$$\forall \alpha, \forall i \in N(f_\alpha) : b_{f_\alpha}(\theta_i) = \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \prod_{j \in N(f_\alpha)/\{i\}} d\theta_j \quad (2.71)$$

- Project the marginal belief to family $\mathcal{F}$:

$$\hat{b}_{f_\alpha}(\theta_i) = \arg \min_{\hat{b}_{f_\alpha}(\theta_i) \in \mathcal{F}} \text{KLD}(b_{f_\alpha}(\theta_i) \| \hat{b}_{f_\alpha}(\theta_i)) \quad (2.72)$$

- Update the message from factors to variables:

$$\mu_{f_\alpha; \theta_i}(\theta_i) = \frac{\hat{b}_{f_\alpha}(\theta_i)}{\mu_{\theta_i; f_\alpha}(\theta_i)} \quad (2.73)$$

- Compute the belief at variable node:

$$\forall i : b_{\theta_i}(\theta_i) \propto \prod_{\alpha \in N(\theta_i)} \mu_{f_\alpha; \theta_i}(\theta_i) \quad (2.74)$$

- Update the message from variable to factor:

$$\forall i, \forall \alpha \in N(\theta_i) : \mu_{\theta_i; f_\alpha}(\theta_i) = \frac{b_{\theta_i}(\theta_i)}{\mu_{f_\alpha; \theta_i}(\theta_i)}. \quad (2.75)$$

At convergence the static point of Expectation Propagation is obtained as a system of equations composed of (2.70)-(2.75). It can be shown that the system of equations composed of (2.70)-(2.75) is equivalent to the system of equations in (2.67). Therefore, Expectation Propagation can be viewed as an iterative method of finding the optimal point of BFE with relaxed moment consistency constraints.

Part I

Linear Estimation

Chapter 3

Vector Approximate Message Passing for Not So Large N.I.I.D. Generalized I/O Linear Models

3.1 Introduction

The input signal recovery in *generalized linear model* (GLM) [29] is a fundamental problem in signal processing. Applications involving this problem include statistics regression [30], wireless communication [31], and machine learning [32]. For instance, within the realm of wireless communication, $p_{\mathbf{y}|\mathbf{x}}$ often characterizes receiver hardware and inference mechanisms, in the symbol detection where $\mathbf{x}$ represents a vector of discrete symbols to be recovered, or in the channel estimation and localization, where $\mathbf{x}$ comprises propagation channel parameters, $\mathbf{A}$ could embody the propagation features, modulation/demodulation schemes, or pilot symbols, depending on the scenario. Viewing this from a Bayesian perspective, the estimation of the random vector $\mathbf{x} \in \mathbb{R}^N$ from observed measurements $\mathbf{y} \in \mathbb{R}^M$ is often tackled through methods such as maximum a posteriori (MAP) estimation and minimum mean square error (MMSE) estimation. However, in certain scenarios, these methods become unviable due to the optimization complexities inherent in MAP or the intractable integrals associated with MMSE.

To tackle this challenge, one established method is approximate inference. Generalized Approximate Message Passing (GAMP) [16] stands out as a prominent and computationally efficient approach within the realm of GLM. Expanding upon the foundation of Approximate Message Passing (AMP) [14] used in standard linear models (SLM), GAMP showcases its prowess in recovering high-dimensional signals. The dynamics of GAMP align with those of AMP, both governed by a state evolution. Nevertheless, GAMP encounters convergence issues when faced with an ill-conditioned measurement matrix.

3.1.1 Prior Work

In response, a solution comes as an extension to the Vector Approximate Message Passing (VAMP) [19] algorithm, tailored for GLM scenarios. This approach, named GVAMP [33], accommodates the challenge of an ill-conditioned matrix. It applies a vector-level Expectation Propagation [7] algorithm on a specially designed factorization scheme. Compared to AMP which requires i.i.d. measurement matrix to converge, VAMP shines particularly for its converging behavior in the scenario of a right rotationally invariant

measurement matrix, with a scalar state evolution having been rigorously established. It is worth noting that GLM-VAMP and its variants often assume a high system dimension, which drives efforts to avoid complex matrix inversions and necessitates additional approximations for efficiency.

However, the original (G)VAMP algorithm solely provides averaged variances. This limitation serves as the impetus for introducing the reVAMP (Revisited- Generalized Vector Approximate Message Passing) [34] for a standard linear model with arbitrary additive Gaussian noise. it gives a Gaussian approximated posterior distribution including inter-parameter correlations.

3.1.2 Main Contribution

In this chapter, reGVAMP (Revisited- Generalized Vector Approximate Message Passing) algorithm is extended from reVAMP to cover GLM. We derive reGVAMP by utilizing the multivariate Gaussian marginalization under EP framework. In reGVAMP, each likelihood and prior is approximated by a Gaussian distribution. Similar to reVAMP, it results in an approximated posterior within the Gaussian family and gives a covariance matrix with inter-parameter correlations.

3.2 Derivation under EP Framework

We consider here a generalized linear model

$$p(\mathbf{x}) = \prod_{i=1}^N p_{x_i}(x_i), \mathbf{z} = \mathbf{A}\mathbf{x}, p_{\mathbf{y}|\mathbf{z}}(\mathbf{z}) = \prod_{j=1}^M p_{y_j|z_j}(z_j), \quad (3.1)$$

where the ratio N/M is a constant. We interpret the linear mixing as a conditional probability

$$p(\mathbf{z}|\mathbf{x}) = \delta(\mathbf{z} - \mathbf{A}\mathbf{x}). \quad (3.2)$$

From this general linear model, a factorization scheme comes up naturally:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \prod_j p_{y_j|z_j}(z_j) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \prod_i p_{x_i}(x_i). \quad (3.3)$$

which we want to approximate by approximated factors (prior and likelihood)

$$q(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \prod_j q_{y_j|z_j}(z_j) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \prod_i q_{x_i}(x_i), \quad (3.4)$$

where q_{x_i} and $q_{y_j|z_j}$ are the Gaussian approximation for prior p_{x_i} and likelihood $p_{y_j|z_j}$. It turns out that EP can be interpreted as an approximate alternating minimization of the Kullback-Leibler divergence (KLD) $KLD(p\|q)$ which becomes:

$$\begin{aligned} & KLD \left[p_{y_j|z_j}(z_j) \prod_{k \neq j} q_{y_k|z_k}(z_k) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \prod_i q_{x_i}(x_i) \right\| \\ & q_{y_j|z_j}^+(z_j) \prod_{k \neq j} q_{y_k|z_k}(z_k) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \prod_i q_{x_i}(x_i) \Big]; \\ & KLD \left[p_{x_i}(x_i) \prod_j q_{y_j|z_j}(z_j) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \prod_{k \neq i} q_{x_k}(x_k) \right\| \\ & q_{x_i}^+(x_i) \prod_j q_{y_j|z_j}(z_j) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \prod_{k \neq i} q_{x_k}(x_k) \Big], \end{aligned} \quad (3.5)$$

where we optimize with respect to approximated likelihood $q_{y_j|z_j}^+$ in the first expression and approximated prior $q_{x_i}^+$ in the second one within Gaussian family. After their optimization

we update the old $q_{y_j|z_j}$ and q_{x_i} with $q_{y_j|z_j}^+$ and $q_{x_i}^+$. We can marginalize out the irrelevant terms when optimizing $q_{y_j|z_j}^+$ and $q_{x_i}^+$ and denote the cavity/extrinsic distribution as

$$\begin{aligned} q_{\mathbf{y}_{\bar{j}}, z_j}(z_j) &\propto \frac{\int q_{\mathbf{y}|z}(\mathbf{z}) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) q_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} d\mathbf{z}_{\bar{j}}}{q_{y_j|z_j}(z_j)}, \\ q_{\mathbf{y}|x_i}(x_i) &\propto \frac{\int q_{\mathbf{x}}(\mathbf{x}) q_{\mathbf{y}|z}(\mathbf{z}) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) d\mathbf{z} d\mathbf{x}_{\bar{i}}}{q_{x_i}(x_i)}. \end{aligned} \quad (3.6)$$

The inner integrals in (3.6) can be considered as joint Gaussian posterior of the output or input with $\mathbf{x}$ or $\mathbf{z}$ marginalized out respectively :

$$\begin{aligned} q_{z|\mathbf{y}}(\mathbf{z}) &= \int q_{\mathbf{y}|z}(\mathbf{z}) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) q_{\mathbf{x}}(\mathbf{x}) d\mathbf{x}; \\ q_{\mathbf{x}|\mathbf{y}}(\mathbf{x}) &= \int q_{\mathbf{x}}(\mathbf{x}) q_{\mathbf{y}|z}(\mathbf{z}) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) d\mathbf{z}. \end{aligned} \quad (3.7)$$

Thus, (3.5) is equivalent to

$$\begin{aligned} KLD[p_{y_j|z_j} \cdot q_{\mathbf{y}_{\bar{j}}, z_j} \| q_{y_j|z_j}^+ \cdot q_{\mathbf{y}_{\bar{j}}, z_j}] &= \mathbb{E} \left[\ln \frac{p_{y_j|z_j}(z_j)}{q_{y_j|z_j}^+(z_j)} \right] \\ KLD[p_{x_i} \cdot q_{\mathbf{y}|x_i} \| q_{x_i}^+ \cdot q_{\mathbf{y}|x_i}] &= \mathbb{E} \left[\ln \frac{p_{x_i}(x_i)}{q_{x_i}^+(x_i)} \right], \end{aligned} \quad (3.8)$$

where the first and second expectations are calculated with respect to the tilted posteriors

$$q_{z|\mathbf{y}}(z_j) \propto p_{y_j|z_j}(z_j) q_{\mathbf{y}_{\bar{j}}, z_j}(z_j) \quad (3.9)$$

$$q_{x_i|\mathbf{y}}(x_i) \propto p_{x_i}(x_i) q_{\mathbf{y}|x_i}(x_i) \quad (3.10)$$

respectively. It is worth noticing that these tilted posteriors can be identified as marginal posterior with Gaussian approximated extrinsic.

3.2.1 Output Node

At the output node, we optimize the first expression in (3.8). For simplicity, we denote

$$\begin{aligned} q_{\mathbf{y}|z}(\mathbf{z}) &= \prod_j q_{y_j|z_j}(z_j) := \mathcal{N}(\mathbf{z} | \mathbf{m}_z, \mathbf{D}_{\tau_z}) \\ q_{\mathbf{x}}(\mathbf{x}) &= \prod_i q_{x_i}(x_i) := \mathcal{N}(\mathbf{x} | \mathbf{m}_x, \mathbf{D}_{\tau_x}), \end{aligned} \quad (3.11)$$

where

$$\mathbf{D}_{\tau_z} = \text{diag}[\tau_z]; \quad \mathbf{D}_{\tau_x} = \text{diag}[\tau_x]. \quad (3.12)$$

The optimization of the first KLD in (3.8) is equivalent to matching the first and second order moment with respect to z_j of the tilted posterior given by (3.9). We first compute the marginalization step of the tilted posterior

$$\begin{aligned} q_{z_j|\mathbf{y}}(z_j) &\propto \int \int \frac{p_{y_j|z_j}(z_j)}{q_{y_j|z_j}(z_j)} q_{\mathbf{y}|z}(\mathbf{z}) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) q_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} d\mathbf{z}_{\bar{j}} \\ &= \frac{p_{y_j|z_j}(z_j)}{q_{y_j|z_j}(z_j)} \int q_{\mathbf{y}|z}(\mathbf{z}) \int \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \mathcal{N}(\mathbf{x} | \mathbf{m}_x, \mathbf{D}_{\tau_x}) d\mathbf{x} d\mathbf{z}_{\bar{j}} \\ &= \frac{p_{y_j|z_j}(z_j)}{q_{y_j|z_j}(z_j)} \int \mathcal{N}(\mathbf{z} | \mathbf{m}_z, \mathbf{D}_{\tau_z}) \mathcal{N}(\mathbf{z} | \mathbf{A}\mathbf{m}_x, \mathbf{A}\mathbf{D}_{\tau_x}\mathbf{A}^T) d\mathbf{z}_{\bar{j}} \end{aligned} \quad (3.13)$$

From the second line to the third line, we consider an auxiliary joint probability

$$\tilde{p}_{\mathbf{x}, \mathbf{z}}(\mathbf{x}, \mathbf{z}) = p_{z|\mathbf{x}}(\mathbf{z} | \mathbf{x}) \tilde{p}_{\mathbf{x}}(\mathbf{x}) := \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \mathcal{N}(\mathbf{x} | \mathbf{m}_x, \mathbf{D}_{\tau_x}). \quad (3.14)$$

Since $\mathbf{z} = \mathbf{A}\mathbf{x}$ and $\mathbf{x} \sim \mathcal{N}(\mathbf{m}_x, \mathbf{D}_{\tau_x})$, we have $\mathbf{z} \sim \mathcal{N}(\mathbf{A}\mathbf{m}_x, \mathbf{A}\mathbf{D}_{\tau_x}\mathbf{A}^T)$. The integral in the last line can be viewed as a multivariate Gaussian marginalization of Gaussian joint posterior (with $\mathbf{x}$ marginalized out)

$$\mathcal{N}(\mathbf{z}|\mathbf{m}_z, \mathbf{D}_{\tau_z})\mathcal{N}(\mathbf{z}|\mathbf{A}\mathbf{m}_x, \mathbf{A}\mathbf{D}_{\tau_x}\mathbf{A}^T) \propto \mathcal{N}(\mathbf{z}|\mathbf{m}_{\hat{z}'}, \mathbf{C}_{\hat{z}'}), \quad (3.15)$$

where

$$\begin{aligned} \mathbf{C}_{\hat{z}'} &= (\mathbf{D}_{\tau_z}^{-1} + (\mathbf{A}\mathbf{D}_{\tau_x}\mathbf{A}^T)^{-1})^{-1} = \mathbf{A}(\mathbf{D}_{\tau_x}^{-1} + \mathbf{A}^T\mathbf{D}_{\tau_z}^{-1}\mathbf{A})^{-1}\mathbf{A}^T; \\ \mathbf{m}_{\hat{z}'} &= \mathbf{C}_{\hat{z}'}(\mathbf{D}_{\tau_z}^{-1}\mathbf{m}_z + (\mathbf{A}\mathbf{D}_{\tau_x}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{m}_x) \\ &= \mathbf{A}(\mathbf{D}_{\tau_x}^{-1} + \mathbf{A}^T\mathbf{D}_{\tau_z}^{-1}\mathbf{A})^{-1}(\mathbf{D}_{\tau_x}^{-1}\mathbf{m}_x + \mathbf{A}^T\mathbf{D}_{\tau_z}^{-1}\mathbf{m}_z). \end{aligned} \quad (3.16)$$

Denote

$$\boldsymbol{\tau}_{\hat{z}'} = \text{diag}[\mathbf{C}_{\hat{z}'}], \quad (3.17)$$

we then obtain the extrinsic/cavity distribution $q_{\mathbf{y}_{\bar{j}}, z_j}(z_j)$ by multivariate Gaussian marginalization

$$\mathcal{N}(z_j|m_{p_j}, \tau_{p_j}) \propto \frac{\int \mathcal{N}(\mathbf{z}|\mathbf{m}_{\hat{z}'}, \mathbf{C}_{\hat{z}'})d\mathbf{z}_{\bar{j}}}{q_{\mathbf{y}_{\bar{j}}, z_j}(z_j)} \propto \frac{\mathcal{N}(z_j|m_{\hat{z}'_j}, \tau_{\hat{z}'_j})}{\mathcal{N}(z_j|m_{z_j}, \tau_{z_j})}, \quad (3.18)$$

where

$$\begin{aligned} \frac{1}{\tau_{p_j}} &= \frac{1}{\tau_{\hat{z}'_j}} - \frac{1}{\tau_{z_j}}; \\ m_{p_j} &= \tau_{p_j} \left(\frac{m_{\hat{z}'_j}}{\tau_{\hat{z}'_j}} - \frac{m_{z_j}}{\tau_{z_j}} \right). \end{aligned} \quad (3.19)$$

Thus, the tilted posterior is proportional to

$$q_{z_j|\mathbf{y}}(z_j) \propto p_{y_j|z_j}(z_j) \mathcal{N}(z_j|m_{p_j}, \tau_{p_j}). \quad (3.20)$$

Matching the first and second order moment of the two parameters in the first KLD in (3.8), we obtain

$$q_{y_j|z_j}^+(z_j) q_{\mathbf{y}_{\bar{j}}, z_j}(z_j) \propto \mathcal{N}(z_j|m_{\hat{z}_j}, \tau_{\hat{z}_j}), \quad (3.21)$$

where $\hat{z}_j$ and τ_{z_j} are calculated as the mean and variances of the tilted posterior

$$\begin{aligned} m_{\hat{z}_j} &= \mathbb{E}_{q_{z_j|\mathbf{y}}}[z_j] = g_z(m_{p_j}, \tau_{p_j}); \\ \tau_{\hat{z}_j} &= \mathbb{E}_{q_{z_j|\mathbf{y}}}[(z_j - m_{\hat{z}_j})^2] = \tau_{p_j} g'_z(m_{p_j}, \tau_{p_j}), \end{aligned} \quad (3.22)$$

where we denote

$$g_z(m_{p_j}, \tau_{p_j}) = \frac{\int z_j p_{y_j|z_j}(z_j) \mathcal{N}(z_j|m_{p_j}, \tau_{p_j}) dz_j}{\int p_{y_j|z_j}(z_j) \mathcal{N}(z_j|m_{p_j}, \tau_{p_j}) dz_j}, \quad (3.23)$$

and the derivative is with respect to the first parameter m_{p_j} .

The new approximated likelihood is then derived as

$$q_{y_j|z_j}^+(z_j) \propto \frac{\mathcal{N}(z_j|m_{\hat{z}_j}, \tau_{\hat{z}_j})}{q_{\mathbf{y}_{\bar{j}}, z_j}(z_j)} = \frac{\mathcal{N}(z_j|m_{\hat{z}_j}, \tau_{\hat{z}_j})}{\mathcal{N}(z_j|m_{p_j}, \tau_{p_j})}. \quad (3.24)$$

It is a quotient of two Gaussian, and thus, the approximated likelihood is updated by

$$q_{y_j|z_j}^+(z_j) = \mathcal{N}(z_j|m_{z_j}^+, \tau_{z_j}^+), \quad (3.25)$$

where

$$\begin{aligned} \frac{1}{\tau_{z_j}^+} &= \frac{1}{\tau_{\hat{z}_j}} - \frac{1}{\tau_{p_j}}; \\ m_{z_j}^+ &= \tau_{z_j}^+ \left(\frac{m_{\hat{z}_j}}{\tau_{\hat{z}_j}} - \frac{m_{p_j}}{\tau_{p_j}} \right). \end{aligned} \quad (3.26)$$

If sequential update is used, the j -th entry of $\mathbf{m}_z$ and τ_z is replaced with $m_{z_j}^+$ and $\tau_{z_j}^+$ before the proceeding to the update of next variable node z_{j+1} . The computation of (3.16) can be simplified as a rank one update. To simplify the computation of the covariance, we define

$$\frac{1}{\Delta_{\tau_{z_j}}} = \frac{1}{\tau_{z_j}^+} - \frac{1}{\tau_{z_j}}. \quad (3.27)$$

The updated covariance according to matrix inverse lemma is

$$\mathbf{C}_{\hat{z}'}^+ = \left(\mathbf{C}_{\hat{z}'}^{-1} + \mathbf{e}_j \Delta_{\tau_{z_j}}^{-1} \mathbf{e}_j^T \right)^{-1} = \mathbf{C}_{\hat{z}'} - \frac{\mathbf{C}_{\hat{z}'} \mathbf{e}_j \mathbf{e}_j^T \mathbf{C}_{\hat{z}'}}{\Delta_{\tau_{z_j}} + \mathbf{e}_j^T \mathbf{C}_{\hat{z}'} \mathbf{e}_j}. \quad (3.28)$$

Similarly, we define

$$\Delta_{m_{z_j}} = \Delta_{\tau_{z_j}} \left(\frac{m_{z_j}^+}{\tau_{z_j}^+} - \frac{m_{z_j}}{\tau_{z_j}} \right), \quad (3.29)$$

and the mean update is obtained as

$$\mathbf{m}_{\hat{z}'}^+ = \mathbf{m}_{\hat{z}'} + \frac{\Delta_{m_{z_j}} - \mathbf{e}_j^T \mathbf{m}_{\hat{z}'}}{\Delta_{\tau_{z_j}} + \mathbf{e}_j^T \mathbf{C}_{\hat{z}'} \mathbf{e}_j} \mathbf{C}_{\hat{z}'} \mathbf{e}_j. \quad (3.30)$$

Thus, the complexity for updating each z_j approximated prior is $O(N^2)$ due to (3.28). The complexity for approximation at the output node is dominated by $O(N^3)$ happening at the start of a sweep where we need to compute $\mathbf{C}_{\hat{z}'}$ according to (3.16) for the first time. Moreover, there are $O(N)$ variable nodes to be updated. In other words, the sweep over vector $\mathbf{z}$ has complexity $O(N^3) + O(N)O(N^2) = O(N^3)$.

3.2.2 Input Node

At the input node, we optimize the second expression in (3.8) with respect to $q_{x_i}^+(x_i)$. The optimization of this KLD is again equivalent to matching the first and second order moment of the tilted posterior since $q_{x_i}^+$ belongs to the Gaussian family.

With the notations defined in (3.11), the tilted posterior is given by

$$\begin{aligned} q_{x_i|\mathbf{y}}(x_i) &\propto \int \int \frac{p_{x_i}(x_i)}{q_{x_i}(x_i)} q_{\mathbf{x}}(\mathbf{x}) q_{\mathbf{y}|\mathbf{z}}(\mathbf{z}) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) d\mathbf{z} d\mathbf{x}_{\bar{i}} \\ &= \frac{p_{x_i}(x_i)}{q_{x_i}(x_i)} \int q_{\mathbf{x}}(\mathbf{x}) \int \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \mathcal{N}(\mathbf{z}|\mathbf{m}_z, \mathbf{D}_{\tau_z}) d\mathbf{z} d\mathbf{x}_{\bar{i}} \\ &= \frac{p_{x_i}(x_i)}{q_{x_i}(x_i)} \int \mathcal{N}(\mathbf{x}|\mathbf{m}_x, \mathbf{D}_{\tau_x}) \mathcal{N}(\mathbf{A}\mathbf{x}|\mathbf{m}_z, \mathbf{D}_{\tau_z}) d\mathbf{x}_{\bar{i}}. \end{aligned} \quad (3.31)$$

Again, the last integral is proportional to a multi-variate Gaussian marginalization of the Gaussian joint posterior (with $\mathbf{z}$ marginalized out), and hence, we compute the Gaussian joint posterior as

$$\mathcal{N}(\mathbf{x}|\mathbf{m}_x, \mathbf{D}_{\tau_x}) \mathcal{N}(\mathbf{A}\mathbf{x}|\mathbf{m}_z, \mathbf{D}_{\tau_z}) \propto \mathcal{N}(\mathbf{x}|\mathbf{m}_{\hat{x}'}, \mathbf{C}_{\hat{x}'}), \quad (3.32)$$

where

$$\begin{aligned} \mathbf{C}_{\hat{x}'} &= (\mathbf{D}_{\tau_x}^{-1} + \mathbf{A}^T \mathbf{D}_{\tau_z}^{-1} \mathbf{A})^{-1}; \\ \mathbf{m}_{\hat{x}'} &= \mathbf{C}_{\hat{x}'} (\mathbf{D}_{\tau_x}^{-1} \mathbf{m}_x + \mathbf{A}^T \mathbf{D}_{\tau_z}^{-1} \mathbf{m}_z). \end{aligned} \quad (3.33)$$

To obtain the marginal distribution, we denote

$$\tau_{\hat{x}'} = \text{diag}[\mathbf{C}_{\hat{x}'}]. \quad (3.34)$$

The extrinsic/cavity distribution $q_{\mathbf{y}|x_i}(x_i)$ can be computed as a quotient of two Gaussian distributions

$$\mathcal{N}(x_i|m_{r_i}, \tau_{r_i}) \propto \frac{\int \mathcal{N}(\mathbf{x}|m_{\hat{\mathbf{x}}'}, \mathbf{C}_{\hat{\mathbf{x}}'}) d\mathbf{x}}{q_{x_i}(x_i)} \propto \frac{\mathcal{N}(x_i|m_{\hat{x}_i'}, \tau_{\hat{x}_i'})}{\mathcal{N}(x_i|m_{x_i}, \tau_{x_i})} \quad (3.35)$$

where

$$\begin{aligned} \frac{1}{\tau_{r_i}} &= \frac{1}{\tau_{\hat{x}_i'}} - \frac{1}{\tau_{x_i}}; \\ m_{r_i} &= \tau_{r_i} \left(\frac{m_{\hat{x}_i'}}{\tau_{\hat{x}_i'}} - \frac{m_{x_i}}{\tau_{x_i}} \right). \end{aligned} \quad (3.36)$$

The tilted posterior is proportional to the product of Gaussian extrinsic and the prior of x_i ,

$$q_{x_i|\mathbf{y}}(x_i) \propto p_{x_i}(x_i) \mathcal{N}(x_i|m_{r_i}, \tau_{r_i}). \quad (3.37)$$

Analog to the discussion (3.21) to (3.23), the approximated posterior is obtained by

$$q_{x_i}^+(x_i) q_{\mathbf{y}|x_i}(x_i) \propto \mathcal{N}(x_i|m_{\hat{x}_i}, \tau_{\hat{x}_i}), \quad (3.38)$$

where

$$\begin{aligned} m_{\hat{x}_i} &= \mathbb{E}_{q_{x_i|\mathbf{y}}}[x_i] = g_x(m_{r_i}, \tau_{r_i}); \\ \tau_{\hat{x}_i} &= \mathbb{E}_{q_{x_i|\mathbf{y}}}[(x_i - m_{\hat{x}_i})^2] = \tau_{r_i} g'_x(m_{r_i}, \tau_{r_i}), \end{aligned} \quad (3.39)$$

with $g_x(m_{r_i}, \tau_{r_i})$ defined by

$$g_x(m_{r_i}, \tau_{r_i}) = \frac{\int x_i p_{x_i}(x_i) \mathcal{N}(x_i|m_{r_i}, \tau_{r_i}) dx_i}{\int p_{x_i}(x_i) \mathcal{N}(x_i|m_{r_i}, \tau_{r_i}) dx_i}, \quad (3.40)$$

and the derivative $g'_x(m_{r_i}, \tau_{r_i})$ is once again with respect to the first parameter m_{r_i} . Hence, the approximated prior can be updated by

$$q_{x_i}^+(x_i) \propto \frac{\mathcal{N}(x_i|m_{\hat{x}_i}, \tau_{\hat{x}_i})}{q_{\mathbf{y}|x_i}(x_i)} = \frac{\mathcal{N}(x_i|m_{\hat{x}_i}, \tau_{\hat{x}_i})}{\mathcal{N}(x_i|m_{r_i}, \tau_{r_i})}. \quad (3.41)$$

This Gaussian quotient can be calculated by

$$q_{x_i}^+(x_i) = \mathcal{N}(x_i|m_{x_i}^+, \tau_{x_i}^+) \propto \frac{\mathcal{N}(x_i|m_{\hat{x}_i}, \tau_{\hat{x}_i})}{\mathcal{N}(x_i|m_{r_i}, \tau_{r_i})}, \quad (3.42)$$

where

$$\begin{aligned} \frac{1}{\tau_{x_i}^+} &= \frac{1}{\tau_{\hat{x}_i}} - \frac{1}{\tau_{r_i}}; \\ m_{x_i}^+ &= \tau_{x_i}^+ \left(\frac{m_{\hat{x}_i}}{\tau_{\hat{x}_i}} - \frac{m_{r_i}}{\tau_{r_i}} \right). \end{aligned} \quad (3.43)$$

We use a similar method as described in (3.27) till (3.30) to simplify the LMMSE step (3.33). Define

$$\frac{1}{\Delta_{\tau_{x_i}}} = \frac{1}{\tau_{x_i}^+} - \frac{1}{\tau_{x_i}}; \quad \Delta_{m_{x_i}} = \Delta_{\tau_{x_i}} \left(\frac{m_{x_i}^+}{\tau_{x_i}^+} - \frac{m_{x_i}}{\tau_{x_i}} \right). \quad (3.44)$$

The update of (3.33) with the new estimated prior is

$$\begin{aligned} \mathbf{C}_{\hat{\mathbf{x}}'}^+ &= \mathbf{C}_{\hat{\mathbf{x}}'} - \frac{\mathbf{C}_{\hat{\mathbf{x}}'} \mathbf{e}_i \mathbf{e}_i^T \mathbf{C}_{\hat{\mathbf{x}}'}}{\Delta_{\tau_{x_i}} + \mathbf{e}_i^T \mathbf{C}_{\hat{\mathbf{x}}'} \mathbf{e}_i}; \\ \mathbf{m}_{\hat{\mathbf{x}}'}^+ &= \mathbf{m}_{\hat{\mathbf{x}}'} + \frac{\Delta_{m_{x_i}} - \mathbf{e}_i^T \mathbf{m}_{\hat{\mathbf{x}}'}}{\Delta_{\tau_{x_i}} + \mathbf{e}_i^T \mathbf{C}_{\hat{\mathbf{x}}'} \mathbf{e}_i} \mathbf{C}_{\hat{\mathbf{x}}'} \mathbf{e}_i. \end{aligned} \quad (3.45)$$

Therefore, at the start of a sweep, the algorithm has a complexity of $O(N^3)$ due to the matrix inverse operation. After that, during the sweep, the algorithm still has a complexity of $O(N^3)$ because there are $O(N)$ messages each of which has a complexity of $O(N^2)$. This algorithm displays a nice symmetry between the input and output nodes. Considering both sweeps for the input and output node, the algorithm has an overall complexity of $O(N^3)$. We conclude the discussion in Algorithm reGVAMP.

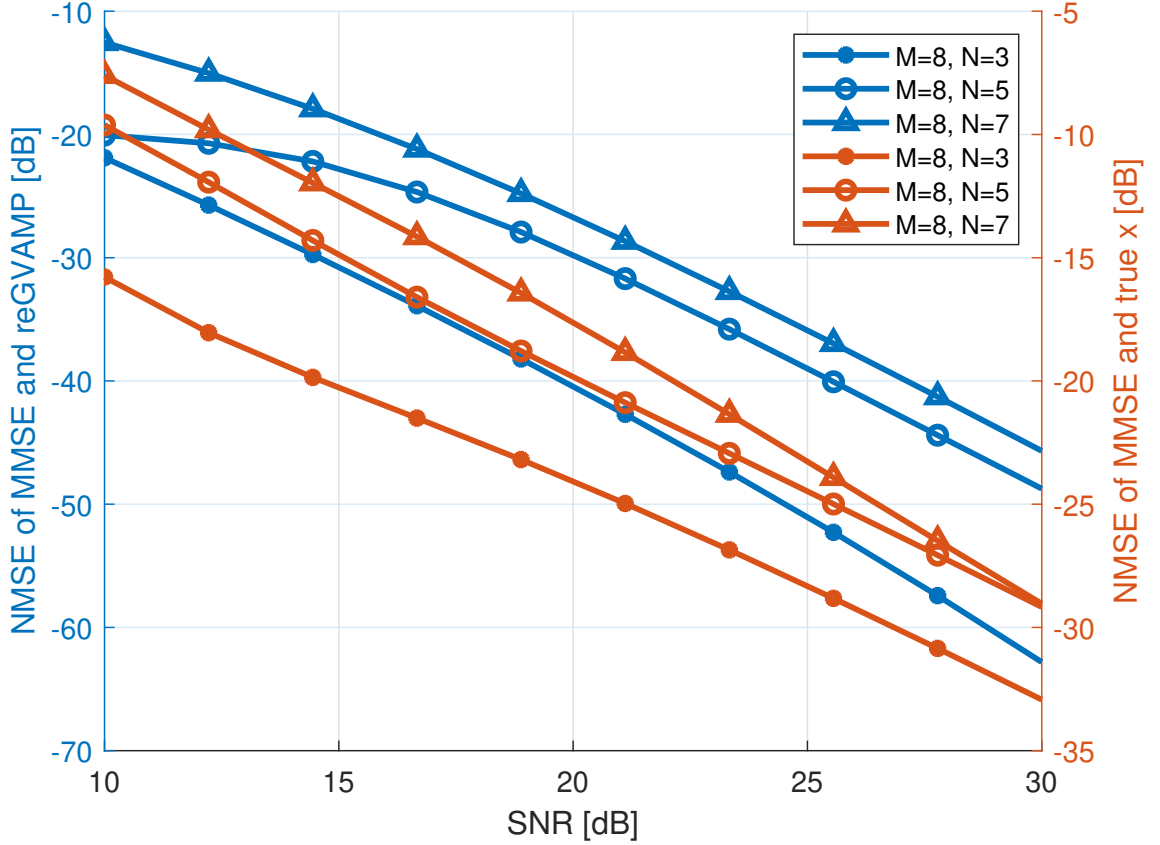


Figure 3.1: Performance of GLM-reVAMP algorithm with different variable length N .

3.3 Simulation Results

We present the results of numerical experiments with varying parameters. In this experiment, we set M to 8 and N to $\{3, 5, 7\}$, respectively. Both the prior distribution $p_{\mathbf{x}}(\mathbf{x})$ and the conditional distribution $p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})$ are assumed to follow mixture Gaussian distributions. Specifically, they can be expressed as:

$$p_{x_i}(x_i) = 0.5\mathcal{N}(x_i|0, \sigma_{x_{1i}}^2) + 0.5\mathcal{N}(x_i|0, \sigma_{x_{2i}}^2), \quad (3.46)$$

$$p_{y_j|z_j}(z_j) = 0.5\mathcal{N}(y_j|z_j, \sigma_{v_{1j}}^2) + 0.5\mathcal{N}(y_j|z_j, \sigma_{v_{2j}}^2), \quad (3.47)$$

where $\sigma_{v_{1j}}$, $\sigma_{v_{2j}}$, $\sigma_{x_{1i}}$ and $\sigma_{x_{2i}}$ are independently and uniformly drawn from the interval $(0, 1]$. The elements of the measurement matrix $\mathbf{A}$ are drawn independently from a Gaussian distribution $\mathcal{N}(0, \gamma)$. By adjusting γ , we can control the signal-to-noise ratio (SNR) of the system. We conducted simulations with SNR values ranging from 10 to 30 dB. Additionally, the recovery performance was evaluated using the normalized mean-squared error (NMSE) defined as $\|\hat{\mathbf{x}} - \mathbf{x}\|^2 / \|\mathbf{x}\|^2$ of one realization. The NMSE of $\hat{\mathbf{x}}_{mmse}$ with respect to $\mathbf{x}$ and $\hat{\mathbf{x}}_{reGVAMP}$ were chosen as performance metrics.

The simulation results are presented in Figure 3.1. It is evident from the figure that as SNR increases, the bias decreases. Furthermore, as N approaches the value of M , the estimation performance deteriorates, although the overall recovery remains satisfactory. These results confirm the effectiveness of the proposed algorithm.

Algorithm 3.1 reGVAMP under EP Framework

Require: $\mathbf{y}, \mathbf{z} = \mathbf{A}\mathbf{x}, p_{\mathbf{x}}(\mathbf{x}), p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z})$

```
1: Initialize:  $\mathbf{m}_{\mathbf{z}}, \boldsymbol{\tau}_{\mathbf{z}}, \mathbf{m}_{\mathbf{x}}, \boldsymbol{\tau}_{\mathbf{x}}$ 
2: repeat[For iteration step  $t$ ]
3:   [Update the Output Nodes]
4:    $\mathbf{C}_{\hat{\mathbf{z}}'} = \mathbf{A}(\mathbf{D}_{\boldsymbol{\tau}_{\mathbf{x}}}^{-1} + \mathbf{A}^T \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{z}}}^{-1} \mathbf{A})^{-1} \mathbf{A}^T$ 
5:    $\mathbf{m}_{\hat{\mathbf{z}}'} = \mathbf{C}_{\hat{\mathbf{z}}'}(\mathbf{D}_{\boldsymbol{\tau}_{\mathbf{z}}}^{-1} \mathbf{m}_{\mathbf{z}} + (\mathbf{A} \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{x}}} \mathbf{A}^T)^{-1} \mathbf{A} \mathbf{m}_{\mathbf{x}})$ 
6:   repeat[For  $j = 1$ ]
7:      $\tau_{\hat{\mathbf{z}}'_j} = [\mathbf{C}_{\hat{\mathbf{z}}'}]_{jj}$ 
8:      $\tau_{p_j} = \frac{\tau_{\hat{\mathbf{z}}'_j} \tau_{\hat{\mathbf{z}}'_j}}{\tau_{\hat{\mathbf{z}}'_j} - \tau_{\hat{\mathbf{z}}'_j}}; m_{p_j} = \frac{m_{\hat{\mathbf{z}}'_j} \tau_{\hat{\mathbf{z}}'_j} - m_{\hat{\mathbf{z}}'_j} \tau_{\hat{\mathbf{z}}'_j}}{\tau_{\hat{\mathbf{z}}'_j} - \tau_{\hat{\mathbf{z}}'_j}}$ 
9:      $m_{\hat{\mathbf{z}}_j} = g_z(m_{p_j}, \tau_{p_j}); \tau_{\hat{\mathbf{z}}_j} = \tau_{p_j} g'_z(m_{p_j}, \tau_{p_j})$ 
10:     $\tau_{z_j}^+ = \frac{\tau_{p_j} \tau_{\hat{\mathbf{z}}_j}}{\tau_{p_j} - \tau_{\hat{\mathbf{z}}_j}}; m_{z_j}^+ = \frac{m_{\hat{\mathbf{z}}_j} \tau_{p_j} - m_{p_j} \tau_{\hat{\mathbf{z}}_j}}{\tau_{p_j} - \tau_{\hat{\mathbf{z}}_j}}$ 
11:    Update  $\mathbf{C}_{\hat{\mathbf{z}}'}^+$  and  $\mathbf{m}_{\hat{\mathbf{z}}'}^+$  according to (3.28), (3.30)
12:  until  $j = M$ 
13:  [Update the Input Nodes]
14:   $\mathbf{C}_{\hat{\mathbf{x}}'} = (\mathbf{D}_{\boldsymbol{\tau}_{\mathbf{x}}}^{-1} + \mathbf{A}^T \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{z}}}^{-1} \mathbf{A})^{-1}$ 
15:   $\mathbf{m}_{\hat{\mathbf{x}}'} = \mathbf{C}_{\hat{\mathbf{x}}'}(\mathbf{D}_{\boldsymbol{\tau}_{\mathbf{x}}}^{-1} \mathbf{m}_{\mathbf{x}} + \mathbf{A}^T \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{z}}}^{-1} \mathbf{m}_{\mathbf{z}})$ 
16:  repeat[For  $i = 1$ ]
17:     $\tau_{\hat{\mathbf{x}}'_i} = [\mathbf{C}_{\hat{\mathbf{x}}'}]_{ii}$ 
18:     $\tau_{r_i} = \frac{\tau_{\hat{\mathbf{x}}'_i} \tau_{\hat{\mathbf{x}}'_i}}{\tau_{\hat{\mathbf{x}}'_i} - \tau_{\hat{\mathbf{x}}'_i}}; m_{r_i} = \frac{m_{\hat{\mathbf{x}}'_i} \tau_{\hat{\mathbf{x}}'_i} - m_{\hat{\mathbf{x}}'_i} \tau_{\hat{\mathbf{x}}'_i}}{\tau_{\hat{\mathbf{x}}'_i} - \tau_{\hat{\mathbf{x}}'_i}}$ 
19:     $m_{\hat{\mathbf{x}}_i} = g_x(m_{r_i}, \tau_{r_i}); \tau_{\hat{\mathbf{x}}_i} = \tau_{r_i} g'_x(m_{r_i}, \tau_{r_i})$ 
20:     $\tau_{x_i}^+ = \frac{\tau_{\hat{\mathbf{x}}_i} \tau_{r_i}}{\tau_{\hat{\mathbf{x}}_i} - \tau_{r_i}}; m_{x_i}^+ = \frac{m_{\hat{\mathbf{x}}_i} \tau_{r_i} - m_{r_i} \tau_{\hat{\mathbf{x}}_i}}{\tau_{\hat{\mathbf{x}}_i} - \tau_{r_i}}$ 
21:    Update  $\mathbf{C}_{\hat{\mathbf{x}}'}^+$  and  $\mathbf{m}_{\hat{\mathbf{x}}'}^+$  according to (3.45)
22:  until  $i = N$ 
23: until Convergence
```

3.4 Concluding Remarks

In this chapter, we proposed an iterative Bayesian estimation method for estimating the GLM input signal which enables us to obtain the posterior mean $\mathbf{m}_{\hat{\mathbf{x}}'}$ and covariance matrix $\mathbf{C}_{\hat{\mathbf{x}}'}$ at the cost of higher complexity. It also gives Gaussian approximation for the prior and likelihood as a byproduct. Further study is required to study its convergence behavior.

Chapter 4

Bethe Free Energy and Extrinsic in Approximate Message Passing

4.1 Introduction

Sparse signal recovery is a fundamental problem in signal processing with a wide range of applications. Many of these problems can be framed as the task of estimating a latent vector $\mathbf{x}$ based on a correlated observation vector $\mathbf{y}$ [6]. In the Bayesian framework, the complexity of Canonical Methods such as MMSE and MAP experiences exponential growth as the dimension of the problem grows.

By exploiting the structure of the models, graphical model based methods prove to be effective. Belief Propagation (BP) transforms the global inference problem into a local inference problem as outlined by [5]. Loopy Belief Propagation (LBP) extends BP by directly employing BP on a factorization scheme for $p(\mathbf{x}|\mathbf{y})$ that may involve loops [8]. In comparison to BP, LBP can be considered as an approximation method.

A limitation of (L)BP is that the (iterative) updating scheme leads to pdfs that correspond to the product of a large number of messages, leading to high complexity. To address this issue, Expectation Propagation (EP) was introduced [35]. EP has been shown to share a similar updating scheme as (L)BP, but for computational efficiency, the messages in (L)BP are projected into a suitable member of the family of exponential distributions [35].

4.1.1 Prior Work

In both [6] and [11], the authors unify EP and BP within the framework of minimizing variational free energy. They demonstrate the close relationship between the fixed points of various message-passing algorithms and the stationary points of Bethe Free Energy (BFE).

EP can serve as an inference method in the linear Gaussian model. However, the computational cost in terms of the message count is quadratic in the data size. Approximate Message Passing (AMP) [12] builds upon EP, but through the application of large system approximations (LSA), it effectively reduces the number of messages to the order of the data size, providing a more computationally efficient approach.

In [18], the authors investigated the fixed points of the Generalized AMP (GAMP) algorithm for generalized linear models (GLMs). They discovered that GAMP shares the same fixed point as the stationary points of the Large System Limit Bethe Free Energy (LSL BFE).

The Component-Wise Conditionally Unbiased (CWCU) Minimum Mean Squared Error (MMSE) estimator is introduced in [24] and rederived in [25] for both joint Gaussian models and linear models. This concept was also used in [36], where the authors call it individual bias compensation. The connection between CWCU MMSE estimation and extrinsic information is explored in [34] specifically for linear Gaussian models.

4.1.2 Main Contributions

Building upon the works of [6] and [37], we present the approximate BFE corresponding to a joint factorization scheme.

We observe that the reGVAMP algorithm, introduced by [37], can be understood as an iterative approach aimed at identifying the stationary points of the proposed BFE. Consequently, this work offers insights into the fixed points of reGVAMP.

The reVAMP method proposed by [34] operates under the assumption of linear Gaussian measurements. In situations where the Gaussian noise is uncorrelated, reVAMP can be considered as a specific instance of reGVAMP.

We also present an alternative derivation of the LSL BFE. Through the application of large system approximations to the stationary points, we substitute certain moment constraints with their equivalent in the large system context. Moreover, the new variance constraints suggest separable approximated posteriors.

We elaborate on the CWCU MMSE discussion, extending it to GLM based on the Gauss-Markov Theorem. This reveals that the extrinsic for both input and output nodes can be interpreted as CWCU MMSE estimation.

4.2 Bethe Free Energy of Generalized Linear Model

In this section, we first give a short introduction to BFE.

4.2.1 Bethe Free Energy

Consider a factorization scheme corresponding to a tree-structured factor graph,

$$p(\mathbf{x}, \mathbf{y}) \propto \prod_{\alpha} f_{\mathbf{x}_{\alpha}}(\mathbf{x}_{\alpha}), \quad (4.1)$$

where $\mathbf{x}_{\alpha}$ is a subvector of $\mathbf{x}$. The tree structure allows an alternative equivalent form [5]

$$p(\mathbf{x}|\mathbf{y}) = \frac{\prod_{\alpha} p(\mathbf{x}_{\alpha})}{\prod_i p(x_i)^{M_i-1}}, \quad (4.2)$$

where M_i is the number of subvectors $\mathbf{x}_{\alpha}$ that contain x_i . In (4.2), the $p(\mathbf{x}_{\alpha})$ and $p(x_i)$ are the exact factor (subvector) marginals and variable marginals.

The concept of variational free energy suggests that to infer the marginals from a tree structured $p(\mathbf{x}, \mathbf{y})$ given in (4.1), we can use as trial distribution

$$q_{\mathbf{x}}(\mathbf{x}) = \frac{\prod_{\alpha} q_{\mathbf{x}_{\alpha}}(\mathbf{x}_{\alpha})}{\prod_i q_{x_i}(x_i)^{M_i-1}}. \quad (4.3)$$

The true marginals can be obtained by [6]

$$\begin{aligned} \min_{q_{\mathbf{x}_\alpha}(\mathbf{x}_\alpha), q_{x_i}(x_i)} F &= D[q(\mathbf{x}) \| \prod_{\alpha} f_{\mathbf{x}_\alpha}(\mathbf{x}_\alpha)]; \\ \text{s.t. } \forall i \forall \alpha, q_{x_i}(x_i) &= \int q_{\mathbf{x}_\alpha}(\mathbf{x}_\alpha) d\mathbf{x}_{\bar{i}}, \end{aligned} \quad (4.4)$$

where we define the shorthand notation (for arbitrary nonnegative functions q, p) $D(q\|p) = \int q(x) \ln \frac{q(x)}{p(x)} dx$ and $\mathbf{x}_{\bar{i}}$ denotes all $\mathbf{x}$ except x_i . The free energy can be expanded as

$$F = \sum_{\alpha} D[q_{\mathbf{x}_\alpha}(\mathbf{x}_\alpha) \| f_{\mathbf{x}_\alpha}(\mathbf{x}_\alpha)] + \sum_i (M_i - 1) H[q_{x_i}(x_i)], \quad (4.5)$$

where $H(\cdot)$ denotes entropy in nats. Note that this representation only holds for a tree structured distribution. For general graphs that contain loops, (4.2) no longer holds. Thus, in cases with loops, (4.5) is only an approximation of the variational free energy. The expression (4.5) is instead called Bethe free energy.

4.2.2 BFE of GLM

We consider a GLM with

$$p(\mathbf{x}) = \prod_{i=1}^N p(x_i), \quad \mathbf{z} = \mathbf{A}\mathbf{x}, \quad p(\mathbf{y}|\mathbf{z}) = \prod_{j=1}^M p(y_j|z_j), \quad (4.6)$$

where the ratio N/M is a constant for large system considerations. We interpret the linear mixing as a conditional probability

$$p(\mathbf{z}|\mathbf{x}) = \delta(\mathbf{z} - \mathbf{A}\mathbf{x}). \quad (4.7)$$

From this general linear model, a joint (loopy) factorization scheme comes up naturally:

$$p(\mathbf{x}, \mathbf{z}|\mathbf{y}) \propto p(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p(\mathbf{y}|\mathbf{z}) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) p(\mathbf{x}). \quad (4.8)$$

According to the definition of BFE (4.5), the associated BFE based on the joint factorization scheme (4.8) is calculated [6] as

$$\begin{aligned} F &= D[q_{\mathbf{x}|\mathbf{y}}(\mathbf{x}) \| p(\mathbf{x})] + D[q_{\mathbf{z}|\mathbf{y}}(\mathbf{z}) \| p(\mathbf{y}|\mathbf{z})] + \sum_i H[q_{x_i|\mathbf{y}}(x_i)] \\ &+ D[b_{\mathbf{x}, \mathbf{z}|\mathbf{y}}(\mathbf{x}, \mathbf{z}) \| \delta(\mathbf{z} - \mathbf{A}\mathbf{x})] + \sum_j H[q_{z_j|\mathbf{y}}(z_j)], \end{aligned} \quad (4.9)$$

where $q_{\mathbf{x}|\mathbf{y}}, q_{\mathbf{z}|\mathbf{y}}, b_{\mathbf{x}, \mathbf{z}|\mathbf{y}}, q_{x_i|\mathbf{y}}$ and $q_{z_j|\mathbf{y}}$ are only approximations of the true posteriors because of the loops in (4.8). Since these approximated posteriors are only locally consistent as is suggested by the constraints in (4.4), they may not correspond to any distribution [5].

As a result, the Bayesian rule can not be used to link $b_{\mathbf{x}, \mathbf{z}|\mathbf{y}}$ with $q_{\mathbf{x}|\mathbf{y}}$ and $q_{\mathbf{z}|\mathbf{y}}$. To use (4.9) as an optimization criterion, we must consider the local consistency between joint and variable marginals as constraints. To make the problem tractable, we use relaxed constraints which contain only the first and second-order moments. To make the discussion concise, define sufficient statistics

$$\phi_{x_i}(x_i) = \begin{bmatrix} x_i \\ x_i^2 \end{bmatrix}; \quad \phi_{z_j}(z_j) = \begin{bmatrix} z_j \\ z_j^2 \end{bmatrix}. \quad (4.10)$$

Reformulate the BFE and the constraints into a Lagrangian function

$$L = F + L_c, \quad (4.11)$$

where L_c is the Lagrange multiplier term

$$\begin{aligned} L_c = & \sum_i \lambda_{x_i}^T \left(\int \phi_{x_i}(x_i) q_{x_i|y}(x_i) dx_i - \int \phi_{x_i}(x_i) q_{x|y}(\mathbf{x}) d\mathbf{x} \right) \\ & + \sum_j \lambda_{z_j}^T \left(\int \phi_{z_j}(z_j) q_{z_j|y}(z_j) dz_j - \int \phi_{z_j}(z_j) q_{z|y}(\mathbf{z}) d\mathbf{z} \right) \\ & + \sum_i \nu_{x_i}^T \left(\int \phi_{x_i}(x_i) q_{x_i|y}(x_i) dx_i - \int \phi_{x_i}(x_i) b_{\mathbf{x},\mathbf{z}|y}(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} \right) \\ & + \sum_j \nu_{z_j}^T \left(\int \phi_{z_j}(z_j) q_{z_j|y}(z_j) dz_j - \int \phi_{z_j}(z_j) b_{\mathbf{x},\mathbf{z}|y}(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} \right). \end{aligned} \quad (4.12)$$

We neglect the normalization constraints to keep the discussion concise. However, one can verify that the Lagrangian multipliers associated with the normalization constraints only act as scaling factors for $b(\cdot)$. Therefore, in the following context, we assume that $b(\cdot)$ are normalized to one.

Since we need to minimize the BFE given by (4.9), the distribution function $b_{\mathbf{x},\mathbf{z}|y}(\mathbf{x}, \mathbf{z})$ must be of the form

$$b_{\mathbf{x},\mathbf{z}|y}(\mathbf{x}, \mathbf{z}) = b_{\mathbf{x}|y}(\mathbf{x}) \delta(\mathbf{z} - \mathbf{A}\mathbf{x}), \quad (4.13)$$

to avoid infinite value of $D[b_{\mathbf{x},\mathbf{z}|y}(\mathbf{x}, \mathbf{z}) \|\delta(\mathbf{z} - \mathbf{A}\mathbf{x})]$, where $b_{\mathbf{x}|y}$ is the function to be optimized. Substitute (4.13) into (4.11) and set the partial derivative of Lagrangian (4.11) with respect to $q_{\mathbf{x}|y}$, $q_{\mathbf{z}|y}$, $b_{\mathbf{x}|y}$, $q_{x_i|y}$ and $q_{z_j|y}$ to zero, we obtain the KKT conditions. Recall the definition of ϕ_{x_i} and ϕ_{z_j} in (4.10). We obtain the Gaussian form by replacing Lagrangian multipliers,

$$q_{\mathbf{x}|y}(\mathbf{x}) \propto p(\mathbf{x}) \mathcal{N}(\mathbf{x} | \mathbf{m}_r, \mathbf{D}_{\tau_r}) \quad (4.14)$$

$$q_{\mathbf{z}|y}(\mathbf{z}) \propto p(\mathbf{z}) \mathcal{N}(\mathbf{z} | \mathbf{m}_p, \mathbf{D}_{\tau_p}) \quad (4.15)$$

$$b_{\mathbf{x}|y} \propto \mathcal{N}(\mathbf{x} | \mathbf{m}_x, \mathbf{D}_{\sigma_x^2}) \mathcal{N}(\mathbf{A}\mathbf{x} | \mathbf{m}_z, \mathbf{D}_{\sigma_z^2}) \quad (4.16)$$

$$\prod_i q_{x_i|y}(x_i) \propto \mathcal{N}(\mathbf{x} | \mathbf{m}_r, \mathbf{D}_{\tau_r}) \mathcal{N}(\mathbf{x} | \mathbf{m}_x, \mathbf{D}_{\sigma_x^2}) \quad (4.17)$$

$$\prod_j q_{z_j|y}(z_j) \propto \mathcal{N}(\mathbf{z} | \mathbf{m}_p, \mathbf{D}_{\tau_p}) \mathcal{N}(\mathbf{z} | \mathbf{m}_z, \mathbf{D}_{\sigma_z^2}), \quad (4.18)$$

where $\mathbf{D}_{\tau_r}$, $\mathbf{D}_{\tau_p}$, $\mathbf{D}_{\sigma_x^2}$ and $\mathbf{D}_{\sigma_z^2}$ are diagonal matrices. These diagonal matrices along with $\mathbf{m}_r$, $\mathbf{m}_p$, $\mathbf{m}_x$ and $\mathbf{m}_z$ correspond to the Lagrange multipliers. Though optimizing the variable marginals may seem like maximizing, they are fully determined by their neighboring factors because of the constraints [11]. Their diagonal elements are denoted by τ_r , τ_p , σ_x^2 and σ_z^2 , respectively. Since the second order moments are linked with variance by $\text{Var}(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2$, using first and second order moments is equivalent to using mean and variance moments.

4.3 Relation to Message Passing and its Stationary Points

These Gaussian distributions can be interpreted as messages. The algorithms reVAMP [34] and reGVAMP [37] can be interpreted as finding the set of consistent messages iteratively in a certain order.

4.3.1 Approximation of Prior $p(\mathbf{x})$

We consider the consistency between (4.14) and (4.17) first. By Gaussian reproduction lemma [12], the product of two Gaussian distributions is still Gaussian. Therefore, (4.17) can also be denoted as

$$\prod_i q_{x_i|\mathbf{y}}(x_i) = \mathcal{N}(\mathbf{x}|\mathbf{m}_{\hat{\mathbf{x}}}, \mathbf{D}_{\hat{\mathbf{x}}}), \quad (4.19)$$

where

$$\mathbf{D}_{\hat{\mathbf{x}}}^{-1} = \mathbf{D}_{\tau_r}^{-1} + \mathbf{D}_{\sigma_x^2}^{-1}; \quad \mathbf{D}_{\hat{\mathbf{x}}}^{-1}\mathbf{m}_{\hat{\mathbf{x}}} = \mathbf{D}_{\tau_r}^{-1}\mathbf{m}_r + \mathbf{D}_{\sigma_x^2}^{-1}\mathbf{m}_x. \quad (4.20)$$

In order to make the pair $(q_{\mathbf{x}|\mathbf{y}}, q_{x_i|\mathbf{y}})$ in (4.14) and (4.17) consistent, we consider $(\mathbf{m}_r, \mathbf{D}_{\tau_r})$ as known and try to derive $(\mathbf{m}_x, \mathbf{D}_{\sigma_x^2})$. Define the Gaussian projection

$$\text{proj}(p) = \arg \min_{q \in \Omega} D_{KL} \left[\frac{p}{Z_p} \| q \right], \quad (4.21)$$

where Ω is the set of uncorrelated Gaussian distributions, Z_p denotes the normalization factor of p and D_{KL} represents Kullback–Leibler (KL) divergence.

The moment consistency implies that

$$\mathcal{N}(\mathbf{x}|\mathbf{m}_x, \mathbf{D}_{\sigma_x^2}) = \frac{\text{proj}[p(\mathbf{x})\mathcal{N}(\mathbf{x}|\mathbf{m}_r, \mathbf{D}_{\tau_r})]}{\mathcal{N}(\mathbf{x}|\mathbf{m}_r, \mathbf{D}_{\tau_r})}. \quad (4.22)$$

This indicates that the message $\mathcal{N}(\mathbf{x}|\mathbf{m}_x, \mathbf{D}_{\sigma_x^2})$ approximates $p(\mathbf{x})$. This update method is the same as updating the message from input node x_i to factor node $\delta(\mathbf{z} - \mathbf{A}\mathbf{x})$ proposed in [37].

Since $p(\mathbf{x})$ is separable, this update scheme contains only scalar integrals.

4.3.2 Extrinsic for Output Node $\mathbf{z}$

Now we need to make $(b_{\mathbf{x},\mathbf{z}|\mathbf{y}}, q_{z_j|\mathbf{y}})$ consistent while assuming (4.16) to be known.

At the stable points, $b_{\mathbf{x}|\mathbf{y}}$, $q_{\mathbf{x}|\mathbf{y}}$ and $\prod_i q_{x_i|\mathbf{y}}(x_i)$ admit the same mean and variance $(\mathbf{m}_{\hat{\mathbf{x}}}, \tau_{\hat{\mathbf{x}}})$. However, their off-diagonal elements may differ. We denote

$$b_{\mathbf{x}|\mathbf{y}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{m}_{\hat{\mathbf{x}}}, \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}), \quad (4.23)$$

where

$$\begin{aligned} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} &= (\mathbf{D}_{\sigma_x^2}^{-1} + \mathbf{A}^T \mathbf{D}_{\sigma_z^2}^{-1} \mathbf{A})^{-1}; \\ \mathbf{m}_{\hat{\mathbf{x}}} &= \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} (\mathbf{D}_{\sigma_x^2}^{-1} \mathbf{m}_x + \mathbf{A}^T \mathbf{D}_{\sigma_z^2}^{-1} \mathbf{m}_z). \end{aligned} \quad (4.24)$$

Likewise, we denote $\prod_j q_{z_j|\mathbf{y}}(z_j)$ as

$$\prod_j q_{z_j|\mathbf{y}}(z_j) = \mathcal{N}(\mathbf{z}|\mathbf{m}_{\hat{\mathbf{z}}}, \mathbf{D}_{\hat{\mathbf{z}}}), \quad (4.25)$$

where

$$\mathbf{D}_{\hat{\mathbf{z}}}^{-1} = \mathbf{D}_{\tau_p}^{-1} + \mathbf{D}_{\sigma_z^2}^{-1}; \quad \mathbf{D}_{\hat{\mathbf{z}}}^{-1}\mathbf{m}_{\hat{\mathbf{z}}} = \mathbf{D}_{\tau_p}^{-1}\mathbf{m}_p + \mathbf{D}_{\sigma_z^2}^{-1}\mathbf{m}_z. \quad (4.26)$$

Since $b_{\mathbf{x},\mathbf{z}|\mathbf{y}}(\mathbf{x}, \mathbf{z}) = b_{\mathbf{x}|\mathbf{y}}(\mathbf{x})\delta(\mathbf{z} - \mathbf{A}\mathbf{x})$. We calculate the marginal distribution of $\mathbf{z}$ as

$$b_{\mathbf{z}|\mathbf{y}}(\mathbf{z}) = \int b_{\mathbf{x},\mathbf{z}|\mathbf{y}}(\mathbf{x}, \mathbf{z}) d\mathbf{x} = \mathcal{N}(\mathbf{z}|\mathbf{A}\mathbf{m}_{\hat{\mathbf{x}}}, \mathbf{A}\mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\mathbf{A}^T) \quad (4.27)$$

We can see that the mean and variances given by $(\mathbf{A}\mathbf{m}_{\hat{\mathbf{x}}}, \text{diag}(\mathbf{A}\mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}\mathbf{A}^T))$ corresponds to the update method for updating the message from $\delta(\mathbf{z} - \mathbf{A}\mathbf{x})$ to $\mathbf{z}$ stated in [37].

Now, look at the variance subsystem. The variance constraints entail

$$\forall k, \mathbf{e}_k^T \mathbf{A} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} \mathbf{A}^T \mathbf{e}_k = \mathbf{e}_k^T \mathbf{D}_{\hat{\mathbf{z}}} \mathbf{e}_k. \quad (4.28)$$

To have a better understanding of the extrinsic of z_k , define

$$\mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}}^{-1} = \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}^{-1} - \frac{1}{\sigma_{z_k}^2} \mathbf{A}_{k,:}^T \mathbf{A}_{k,:}, \quad (4.29)$$

where $\mathbf{A}_{k,:}$ denotes the k -th row of matrix $\mathbf{A}$. Applying the matrix inversion lemma, the LHS of (4.28) becomes

$$\mathbf{A}_{k,:} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}}} \mathbf{A}_{k,:}^T = \frac{\sigma_{z_k}^2 \mathbf{A}_{k,:} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}} \mathbf{A}_{k,:}^T}{\sigma_{z_k}^2 + \mathbf{A}_{k,:} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}} \mathbf{A}_{k,:}^T}. \quad (4.30)$$

Substituting (4.26), (4.30) into (4.28) yields

$$\tau_{p_k} = \mathbf{A}_{k,:} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}} \mathbf{A}_{k,:}^T. \quad (4.31)$$

Since $\mathbf{A}_{k,:}$ is independent of $\mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}}$, in the LSL, we get [38]

$$\mathbf{A}_{k,:} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}} \mathbf{A}_{k,:}^T \simeq \text{tr}[\mathbf{\Theta}_k \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}], \quad (4.32)$$

where $\mathbf{\Theta}_k = \mathbb{E}[\mathbf{A}_{k,:}^T \mathbf{A}_{k,:}]$.

If we further assume each entry of $\mathbf{A}$ to have deterministic absolute value but i.i.d. signs, it follows that $\mathbf{\Theta}_k = \text{diag}(\mathbf{S}_{k,:})$, where $\mathbf{S} = \mathbf{A} \cdot \mathbf{A}$ denotes the element-wise square of $\mathbf{A}$. This further simplifies (4.32)

$$\tau_{p_k} \simeq \text{tr}[\mathbf{\Theta}_k \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}}}] = \mathbf{S}_{k,:} \cdot \boldsymbol{\tau}_{\hat{\mathbf{x}}}. \quad (4.33)$$

A similar analysis can be done for the mean subsystem. Now we assume that the variance has been made consistent. The consistency of the mean implies that

$$\mathbf{e}_k^T \mathbf{A} \mathbf{m}_{\hat{\mathbf{x}}} = \mathbf{e}_k^T \mathbf{m}_{\hat{\mathbf{z}}}. \quad (4.34)$$

Denote

$$\mathbf{n}_{\hat{\mathbf{x}}} = \mathbf{D}_{\sigma_{\hat{\mathbf{x}}}^2}^{-1} \mathbf{m}_{\mathbf{x}} + \mathbf{A}^T \mathbf{D}_{\sigma_{\hat{\mathbf{z}}}^2}^{-1} \mathbf{m}_{\mathbf{z}}; \quad \mathbf{n}_{\hat{\mathbf{x}},\bar{k}} = \mathbf{n}_{\hat{\mathbf{x}}} - \frac{m_{z_k}}{\sigma_{z_k}^2} \mathbf{A}_{k,:}^T. \quad (4.35)$$

By applying the matrix inversion lemma, we can rewrite the expression for the k^{th} element of $\mathbf{A} \mathbf{m}_{\hat{\mathbf{x}}}$ in (4.27)

$$\begin{aligned} \mathbf{A}_{k,:} \mathbf{m}_{\hat{\mathbf{x}}} &= \frac{\sigma_{z_k}^2}{\mathbf{A}_{k,:} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}} \mathbf{A}_{k,:}^T + \sigma_{z_k}^2} \mathbf{A}_{k,:} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}} \mathbf{n}_{\hat{\mathbf{x}},\bar{k}} \\ &+ \frac{\mathbf{A}_{k,:} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}} \mathbf{A}_{k,:}^T}{\mathbf{A}_{k,:} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}} \mathbf{A}_{k,:}^T + \sigma_{z_k}^2} m_{z_k} \end{aligned} \quad (4.36)$$

Substitute (4.31) into (4.36) and equate $\mathbf{A}_{k,:} \mathbf{m}_{\hat{\mathbf{x}}}$ with $m_{\hat{z}_k}$ given by (4.26) to obtain the extrinsic mean

$$m_{p_k} = \mathbf{A}_{k,:} \mathbf{C}_{\hat{\mathbf{x}}\hat{\mathbf{x}},\bar{k}} \mathbf{n}_{\hat{\mathbf{x}},\bar{k}}. \quad (4.37)$$

We can also calculate the extrinsic mean as a function of $\mathbf{m}_{\hat{\mathbf{x}}}$.

By combining (4.26) and (4.34), we see

$$m_{p_k} = \left(1 + \frac{\tau_{p_k}}{\sigma_{z_k}^2}\right) \mathbf{A}_{k,:} \mathbf{m}_{\hat{\mathbf{x}}} - \frac{\tau_{p_k}}{\sigma_{z_k}^2} m_{z_k}. \quad (4.38)$$

The discussions above only hold at the stable point where $b_{\mathbf{x}|\mathbf{y}}$ has the same mean and variance with $\prod_i q_{x_i|\mathbf{y}}(x_i)$. Therefore, we can view the relations given by (4.33) and (4.34) as alternative constraints.

4.3.3 Approximation of Likelihood $p(\mathbf{y}|\mathbf{z})$

At this point, we consider the extrinsic $\mathcal{N}(\mathbf{z}|\mathbf{m}_p, \mathbf{D}_{\tau_p})$ to be given. To make (4.15) and (4.18) consistent, we use similar methods described in (4.19) till (4.22), which gives

$$\mathcal{N}(\mathbf{z}|\mathbf{m}_z, \mathbf{D}_{\sigma_z^2}) = \frac{\text{proj}[p(\mathbf{y}|\mathbf{z})\mathcal{N}(\mathbf{z}|\mathbf{m}_p, \mathbf{D}_{\tau_p})]}{\mathcal{N}(\mathbf{z}|\mathbf{m}_p, \mathbf{D}_{\tau_p})}. \quad (4.39)$$

This separable update method indicates that the distribution $\mathcal{N}(\mathbf{z}|\mathbf{m}_z, \mathbf{D}_{\sigma_z^2})$ stands for the approximate likelihood. In [37], the update of messages from $\mathbf{z}$ to $\delta(\mathbf{z} - \mathbf{A}\mathbf{x})$ employs the same method as outlined in (4.39).

4.4 Derivation of LSL-BFE

Observe the stable point relation (4.33) and (4.34) which are alternative constraints for making the pairs $(b_{\mathbf{x},\mathbf{z}|\mathbf{y}}, q_{z_j|\mathbf{y}})$ consistent. By using this alternative constraint, we modify the last Lagrangian term of $\mathbb{E}_{q_{z_j|\mathbf{y}}}[\phi_{z_j}(z_j)] = \mathbb{E}_{b_{\mathbf{x},\mathbf{z}|\mathbf{y}}}[\phi_{z_j}(z_j)]$ to

$$\begin{aligned} & \sum_j u_{z_j, \text{mean}} \left(m_{\hat{z}_j} - \sum_i \mathbf{A}_{ji} \int x_i b_{\mathbf{x}|\mathbf{y}}(\mathbf{x}) d\mathbf{x} \right) \\ & + \sum_j u_{z_j, \text{var}} \left(\tau_{p_j} - \sum_i \mathbf{S}_{ji} \text{Var}_{b_{\mathbf{x}|\mathbf{y}}}(x_i) \right) \end{aligned} \quad (4.40)$$

With this replacement, we see that $b_{\mathbf{x}|\mathbf{y}}(\mathbf{x})$ is now separable by considering the variational derivative of (4.11). Furthermore, by combining (4.13), the consistency between separable $b_{\mathbf{x}|\mathbf{y}}$ and $\forall i, q_{x_i|\mathbf{y}}(x_i)$ implies that the following two terms in (4.9) are identical

$$D[b_{\mathbf{x},\mathbf{z}|\mathbf{y}}(\mathbf{x}, \mathbf{z}) \|\delta(\mathbf{z} - \mathbf{A}\mathbf{x})] + \sum_i H[q_{x_i|\mathbf{y}}(x_i)] = 0 \quad (4.41)$$

Now we will consider the relation between $\mathbf{z}$ side and (4.40). The constraints given by (4.40) are applied to the posterior mean and extrinsic variance of node $\mathbf{z}$.

We use the ansatz that $q_{\mathbf{z}|\mathbf{y}}$ is separable. Indeed, in (4.9), if we look at the derivative with respect to $q_{\mathbf{z}|\mathbf{y}}$, the term $D[q_{\mathbf{z}|\mathbf{y}}(\mathbf{z}) \| p(\mathbf{y}|\mathbf{z})]$ implies a separable $q_{\mathbf{z}|\mathbf{y}}$. Furthermore, the constraints (4.40) also indicate a separable extrinsic for $q_{\mathbf{z}|\mathbf{y}}$. Therefore, in large system limit, we can use strict marginal constraint for the pairs $(q_{\mathbf{z}|\mathbf{y}}, q_{z_j|\mathbf{y}})$, which entails $q_{\mathbf{z}|\mathbf{y}} = \prod_j q_{z_j|\mathbf{y}}(z_j)$. This leads to a hybrid message passing algorithm [6].

Calculate the derivative of the Lagrangian function with respect to $q_{\mathbf{z}|\mathbf{y}}$ for the BFE along with the posterior constraint given by (4.40)

$$\begin{aligned} & \frac{d}{d q_{\mathbf{z}|\mathbf{y}}(\mathbf{z})} D[q_{\mathbf{z}|\mathbf{y}}(\mathbf{z}) \| p(\mathbf{y}|\mathbf{z})] + H[q_{\mathbf{z}|\mathbf{y}}(\mathbf{z})] + \sum_j u_{z_j, \text{mean}} m_{\hat{z}_j} \\ & = -\log[p(\mathbf{y}|\mathbf{z})] + \mathbf{u}_{z_j, \text{mean}}^T \mathbf{z} + c \end{aligned} \quad (4.42)$$

Combining with the definition of τ_p in (4.15), the extrinsic constraint in (4.40) suggests that $q_{\mathbf{z}|\mathbf{y}}$ must be of the form

$$q_{\mathbf{z}|\mathbf{y}} \propto p(\mathbf{y}|\mathbf{z}) \mathcal{N}(\mathbf{z}|\mathbf{m}_p, \mathbf{D}_{\tau_p}), \quad (4.43)$$

where $\mathbf{D}_{\tau_p} = \text{diag}(\mathbf{S}\tau_{\hat{x}})$.

Recall the variation derivative rule

$$\frac{d}{dp(x)} \int p(y) \log \frac{p(y)}{q(y)} dy = \log[p(x)] - \log[q(x)] + 1 \quad (4.44)$$

It indicates that we can modify the extrinsic for $\mathbf{z}$ additively by adding terms of the form $D(q_{\mathbf{z}|\mathbf{y}}\|q_{\mathbf{z}}^e)$.

Assume

$$q_{\mathbf{z}}^e(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z, \mathbf{D}_{\tau_p}). \quad (4.45)$$

To satisfy the implicit extrinsic variance constraint given by (4.40) in the Lagrangian function explicitly, the objective function (which is LSL BFE) is equivalent to

$$\begin{aligned} F_{LSL} &= D[q_{\mathbf{x}|\mathbf{y}}(\mathbf{x})\|p(\mathbf{x})] + D[q_{\mathbf{z}|\mathbf{y}}(\mathbf{z})\|p(\mathbf{y}|\mathbf{z})] \\ &+ H[q_{\mathbf{z}|\mathbf{y}}(\mathbf{z})] + D(q_{\mathbf{z}|\mathbf{y}}\|q_{\mathbf{z}}^e). \end{aligned} \quad (4.46)$$

Furthermore, because of the introduction of auxiliary variable τ_p , we also need to minimize BFE with respect to it.

As $\boldsymbol{\mu}_z$ is an unconstrained free variable, we optimize it directly by zeroing the derivative concerning it. Expand the terms $H[q_{\mathbf{z}|\mathbf{y}}(\mathbf{z})] + D(q_{\mathbf{z}|\mathbf{y}}\|q_{\mathbf{z}}^e)$ in (4.46)

$$\begin{aligned} &H[q_{\mathbf{z}|\mathbf{y}}(\mathbf{z})] + D(q_{\mathbf{z}|\mathbf{y}}\|q_{\mathbf{z}}^e) \\ &= c + (\mathbf{m}_{\hat{z}} - \boldsymbol{\mu}_p)^T \mathbf{D}_{\tau_p}^{-1} (\mathbf{m}_{\hat{z}} - \boldsymbol{\mu}_z). \end{aligned} \quad (4.47)$$

We see the minimal is achieve at $\boldsymbol{\mu}_z = \mathbf{m}_{\hat{z}}$.

4.5 Relation to CWCU MMSE Estimator

The algorithm proposed by [34] can be interpreted as an iterative method of finding consistent messages in (4.14) - (4.18) in the cases where $p(\mathbf{y}|\mathbf{z})$ is modeled as AWGN channel. [34] also shows the close relation between CWCU LMMSE estimation and the extrinsic. In the following, we will interpret the extrinsic as CWCU LMMSE estimation based on the Gauss-Markov theorem.

Based on the discussion of the previous section, when deriving the extrinsic for $\mathbf{z}$ and $\mathbf{x}$, we find the system to be equivalent to a Gaussian linear model. Therefore, we can use the approximate prior and approximate likelihood as if they are the true prior and likelihood when deriving the extrinsics without large system approximations [25].

Consider jointly Gaussian $\mathbf{y}$ and x (scalar)

$$\begin{bmatrix} \mathbf{y} \\ x \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{m}_y \\ m_x \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{yy} & \mathbf{C}_{yx} \\ \mathbf{C}_{xy} & C_{xx} \end{bmatrix} \right) \quad (4.48)$$

Then the extrinsic $p(\mathbf{y}|x)$ is Gaussian and based on Gaussi-Markov theorem

$$\begin{aligned} -2 \ln p(\mathbf{y}|x) &= c + (\mathbf{y} - \mathbf{m}_{y|x})^T \mathbf{C}_{y|x}^{-1} (\mathbf{y} - \mathbf{m}_{y|x}), \text{ with} \\ \mathbf{m}_{y|x} &= \mathbf{m}_y + \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} (\mathbf{x} - m_x), \\ \mathbf{C}_{y|x} &= \mathbf{C}_{yy} - \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \end{aligned} \quad (4.49)$$

Interpreting (4.49) as a pdf in x (which Fisher called fiducial statistics), we can rewrite this quadratic exponent as

$$\begin{aligned}
-2 \ln p(\mathbf{y}|x) &= c(\mathbf{y}) + (x - \hat{x}_{CL})^2 / \mathbf{C}_{\tilde{x}_{CL}\tilde{x}_{CL}}, \\
\hat{x}_{CL} &= m_x + d \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} (\mathbf{y} - \mathbf{m}_y) = d \hat{x}_L + (1 - d) m_x \\
\mathbf{C}_{\tilde{x}_{CL}\tilde{x}_{CL}} &= d \mathbf{C}_{\tilde{x}_L\tilde{x}_L}, \\
&\text{with} \\
\hat{x}_L &= m_x + \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} (\mathbf{y} - \mathbf{m}_y), \quad \mathbf{C}_{\tilde{x}_L\tilde{x}_L} = \mathbf{C}_{xx} - \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \\
d &= \frac{\mathbf{C}_{xx}}{\mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx}} \geq 1,
\end{aligned} \tag{4.50}$$

where $\hat{x}_{CL}$, $\mathbf{C}_{\tilde{x}_{CL}\tilde{x}_{CL}}$ are the CWCU LMMSE estimate and error variance, and $\hat{x}_L$, $\mathbf{C}_{\tilde{x}_L\tilde{x}_L}$ are the LMMSE (and hence MMSE since Gaussian) estimate and error variance.

Now we will investigate the vector case. Define the operation $\text{Diag}(\mathbf{C}) = \text{diag}[\text{diag}(\mathbf{C})]$, which returns a diagonal matrix composed of the diagonal elements of $\mathbf{C}$.

Interpreting the previous x as a component x_i of a vector $\mathbf{x}$, we can write

$$\begin{aligned}
\hat{\mathbf{x}}_{CL} &= \mathbf{m}_x + \mathbf{D} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} (\mathbf{y} - \mathbf{m}_y) = \mathbf{D} \hat{\mathbf{x}}_L + (\mathbf{I} - \mathbf{D}) \mathbf{m}_x \\
\mathbf{C}_{\tilde{\mathbf{x}}_{CL}\tilde{\mathbf{x}}_{CL}} &= \mathbf{C}_{\tilde{\mathbf{x}}_L\tilde{\mathbf{x}}_L} + (\mathbf{D} - \mathbf{I}) \mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L} (\mathbf{D} - \mathbf{I}) \\
&\text{with} \\
\mathbf{D} &= \text{Diag}(\mathbf{C}_{xx}) [\text{Diag}(\mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L})]^{-1}, \quad \mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L} = \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx}
\end{aligned} \tag{4.51}$$

where the expression for $\mathbf{C}_{\tilde{\mathbf{x}}_{CL}\tilde{\mathbf{x}}_{CL}}$ follows from $\tilde{\mathbf{x}}_{CL} = \mathbf{x} - \hat{\mathbf{x}}_{CL} = \tilde{\mathbf{x}}_L - (\mathbf{D} - \mathbf{I}) \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} (\mathbf{y} - \mathbf{m}_y)$ and the two terms in this difference are decorrelated by the orthogonality property of LMMSE estimation.

Next, we'll show: $\mathbf{D} = \text{diag}(\boldsymbol{\tau}_{CL} / \boldsymbol{\tau}_L)$, where $\boldsymbol{\tau}_L = \text{diag}(\mathbf{C}_{\tilde{\mathbf{x}}_L\tilde{\mathbf{x}}_L})$ and $\boldsymbol{\tau}_{CL} = \text{diag}(\mathbf{C}_{\tilde{\mathbf{x}}_{CL}\tilde{\mathbf{x}}_{CL}})$.

$$\begin{aligned}
\mathbf{C}_{\tilde{\mathbf{x}}_{CL}\tilde{\mathbf{x}}_{CL}} &= \mathbf{C}_{\tilde{\mathbf{x}}_L\tilde{\mathbf{x}}_L} + (\mathbf{D} - \mathbf{I}) \mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L} (\mathbf{D} - \mathbf{I}) \\
&= \mathbf{C}_{xx} - \mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L} \mathbf{D} - \mathbf{D} \mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L} + \mathbf{D} \mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L} \mathbf{D}
\end{aligned} \tag{4.52}$$

Calculate the diagonal elements

$$\begin{aligned}
\text{diag}(\boldsymbol{\tau}_{CL}) &= \text{Diag}(\mathbf{C}_{\tilde{\mathbf{x}}_{CL}\tilde{\mathbf{x}}_{CL}}) = \text{Diag}(\mathbf{C}_{xx}) \\
&+ \mathbf{D} \text{Diag}(\mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L}) \mathbf{D} - \text{Diag}(\mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L}) \mathbf{D} - \mathbf{D} \text{Diag}(\mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L}) \\
&= \text{Diag}(\mathbf{C}_{xx}) [\text{Diag}(\mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L})]^{-1} \text{Diag}(\mathbf{C}_{xx}) - \text{Diag}(\mathbf{C}_{xx}),
\end{aligned} \tag{4.53}$$

where we use $\mathbf{D} = \text{Diag}(\mathbf{C}_{xx}) [\text{Diag}(\mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L})]^{-1}$ in (4.51).

Now we want to show $\mathbf{D} \text{diag}(\boldsymbol{\tau}_L) = \text{diag}(\boldsymbol{\tau}_{CL})$:

$$\begin{aligned}
\mathbf{D} \text{diag}(\boldsymbol{\tau}_L) &= \mathbf{D} \text{Diag}(\mathbf{C}_{\tilde{\mathbf{x}}_L\tilde{\mathbf{x}}_L}) \\
&= \text{Diag}(\mathbf{C}_{xx}) [\text{Diag}(\mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L})]^{-1} \\
&\cdot [\text{Diag}(\mathbf{C}_{xx}) - \text{Diag}(\mathbf{C}_{\hat{\mathbf{x}}_L\hat{\mathbf{x}}_L})] = \text{diag}(\boldsymbol{\tau}_{CL})
\end{aligned} \tag{4.54}$$

The extrinsic for $\mathbf{x}$ without large system approximations can be interpreted as CWCU MMSE estimation from the Gaussian model

$$\begin{bmatrix} \mathbf{m}_z \\ \mathbf{x} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{A} \mathbf{m}_x \\ \mathbf{m}_x \end{bmatrix}, \begin{bmatrix} \mathbf{A} \mathbf{D}_{\sigma_x^2} \mathbf{A}^T + \mathbf{D}_{\sigma_z^2} & \mathbf{A} \mathbf{D}_{\sigma_x^2} \\ \mathbf{D}_{\sigma_x^2} \mathbf{A}^T & \mathbf{D}_{\sigma_x^2} \end{bmatrix} \right). \tag{4.55}$$

The underlying equivalent Gaussian linear model is

$$\mathbf{m}_z = \mathbf{A} \mathbf{x} + \mathbf{v}_x \tag{4.56}$$

where $\mathbf{x} \sim \mathcal{N}(\mathbf{m}_x, \mathbf{D}_{\sigma_x^2})$ and $\mathbf{v}_x \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_{\sigma_z^2})$.

Likewise, we can interpret the extrinsic for $\mathbf{z}$ as CWCU MMSE estimation from

$$\begin{bmatrix} \mathbf{A}\mathbf{m}_x \\ \mathbf{z} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{m}_z \\ \mathbf{m}_z \end{bmatrix}, \begin{bmatrix} \mathbf{D}_{\sigma_z^2} + \mathbf{A}\mathbf{D}_{\sigma_x^2}\mathbf{A}^T & \mathbf{D}_{\sigma_z^2} \\ \mathbf{D}_{\sigma_z^2} & \mathbf{D}_{\sigma_z^2} \end{bmatrix} \right). \quad (4.57)$$

The underlying equivalent Gaussian linear model is

$$\mathbf{A}\mathbf{m}_x = \mathbf{z} + \mathbf{v}_z \quad (4.58)$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{m}_z, \mathbf{D}_{\sigma_z^2})$ and $\mathbf{v}_z \sim \mathcal{N}(\mathbf{0}, \mathbf{A}\mathbf{D}_{\sigma_x^2}\mathbf{A}^T)$.

4.6 Concluding Remarks

In this chapter, we studied the BFE of GLMs using a joint factorization scheme. This factorization allows us to extract approximate priors and likelihood. By looking at the stationary point in LSL we replace the non-separable constraints with separable ones. This leads to the LSL BFE. This chapter also interprets extrinsics for both input and output nodes as CWCU LMMSE estimation operations.

Chapter 5

Towards Convergent Approximate Message Passing by Alternating Constrained Minimization of Bethe Free Energy

5.1 Introduction

In the Gaussian noise case, a sparse signal vector $\mathbf{x}$ can be recovered using the signal model: $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}$, where $\mathbf{y}$ is the observed data, $\mathbf{A}$ is the known measurement or sensing matrix of dimension $M \times N$, typically with $M < N$. In the sparse model case, $\mathbf{x}$ contains only K non-zero entries, with $K < M < N$. Sparse Bayesian Learning (SBL) is a Bayesian inference algorithm proposed by [39] and [40]. SBL is based on a hierarchical prior on the sparse coefficients $\mathbf{x}$, inducing sparsity by choosing priors for the hyperparameters that make a portion of the estimate $\hat{\mathbf{x}}$ zero. The Linear Minimum Mean Squared Error (LMMSE) estimation step in SBL at each iteration involves matrix inversion, which makes it computationally complex [41].

The Approximate Message Passing (AMP) algorithm has been introduced to reduce the complexity of Belief Propagation, from $2MN$ to $M + N$ messages. Generalized AMP (GAMP) allows non-Gaussian priors and measurement processes. But convergence of (G)AMP can be problematic for some matrices $\mathbf{A}$. Existing converging AMP versions introduced so far: 1) adding the Alternating Direction Method of Multipliers (ADMM) [42] leading to a higher complexity ADMM-GAMP, 2) exploiting part of the singular value decomposition (SVD) of the measurement matrix in Vector AMP (VAMP) [19], [33] or esp. Unitarily Transformed UT-AMP [43] (but which do not allow to handle n.i.i.d. priors conveniently), 3) introducing damping [18], but with typically difficult to determine damping requirements.

5.1.1 Contributions

- We propose a convergent version of GAMP, AMBGAMP, which applies alternating minimization to an augmented Lagrangian of a large system limit of the Bethe free Energy (BFE). AMBGAMP can be interpreted as applying a simplified ADMM to the BFE, with a constrained Lagrange multiplier parameterization for the mean constraint, and a quadratic optimization subproblem being solved by a gradient update with line search.

The ADMM is complemented with a fixed point iteration for the variance constraint.

- We show that AMBGAMP converges to the LMMSE estimate in the Gaussian case.
- We provide a convergence analysis of the variance subsystem.
- We show that in the Gaussian case the variances converge to the optimal MSE values in the large system limit.
- We provide a convergence analysis of the mean subsystem.

5.2 System Model

The data model considered in GAMP is essentially a linear mixing model represented by

$$\mathbf{z} = \mathbf{A} \mathbf{x}, p_{\mathbf{x}}(\mathbf{x}), p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z}) \quad (5.1)$$

with n.i.i.d. prior $p_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^N p_{x_i}(x_i)$ and n.i.i.d. noisy measurements $p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z}) = \prod_{k=1}^M p_{y_k|z_k}(y_k|z_k)$. In the case of Gaussian measurement noise, we have $\mathbf{y} = \mathbf{z} + \mathbf{v}$ with independent zero-mean n.i.i.d. Gaussian noise $\mathbf{v}$ with variance vector $\boldsymbol{\sigma}_v^2 = [\sigma_{v1}^2, \dots, \sigma_{vM}^2]^T$. We shall also consider the case of a zero mean Gaussian prior for $\mathbf{x}$ with variances denoted as $\sigma_{x_i}^2$. We represent the vector $\boldsymbol{\sigma}_x^2 = [\sigma_{x1}^2, \dots, \sigma_{xN}^2]^T$. In Bayesian estimation, we are interested in the posterior, which is given by

$$p_{\mathbf{x},\mathbf{z}|\mathbf{y}}(\mathbf{x}, \mathbf{z}|\mathbf{y}) = \frac{e^{-\sum_{i=1}^N f_{x_i}(x_i) - \sum_{k=1}^M f_{z_k}(z_k)}}{Z(\mathbf{y})} \mathbb{1}_{\{\mathbf{z}=\mathbf{A}\mathbf{x}\}}, \quad (5.2)$$

with the negative log-likelihoods defined as $f_{x_i}(x_i) = -\ln p_{x_i}(x_i)$, $f_{z_k}(z_k) = -\ln p_{y_k|z_k}(y_k|z_k)$, where the equality in case of $f_{z_k}(z_k)$ is up to constants that may depend on $\mathbf{y}$ (and which are absorbed in the normalization constant $Z(\mathbf{y})$). The problem in Bayesian estimation is the computation of this constant $Z(\mathbf{y})$ and of the posterior means and variances. Belief propagation is a message passing technique that allows to compute the posterior marginals. However, due to loops in the factor graph, loopy belief propagation may have convergence issues and is furthermore still relatively complex. GAMP is an approximate belief propagation technique which is motivated by asymptotic considerations in which the rows and columns of the measurement matrix $\mathbf{A}$ are considered as random and independent, in which case GAMP can actually produce the correct posterior marginals. In any case, GAMP computes a separable approximate posterior of the form

$$q_{\mathbf{x},\mathbf{z}}(\mathbf{x}, \mathbf{z}) = q_{\mathbf{x}}(\mathbf{x}) q_{\mathbf{z}}(\mathbf{z}) = \prod_{i=1}^N q_{x_i}(x_i) \prod_{k=1}^M q_{z_k}(z_k), \quad (5.3)$$

in which the dependence on $\mathbf{y}$ has been omitted. The GAMP algorithm [16], [18] appears in the table for Algorithm 1. We only consider here Sum-Product GAMP (for MMSE estimation, as opposed to Max-Sum GAMP for MAP estimation).

5.3 Proposed AMBGAMP

The abbreviation AMB stands for ACM-LSL-BFE, which stands for Alternating Constrained Minimization of the LSL of the BFE. AMBGAMP employs most of the same updates as GAMP, but GAMP does not apply a strict alternating minimization (block coordinate descent) principle, particularly in the presence of constraints. Previous work [17] has demonstrated that any fixed point of the GAMP algorithm is a critical point of the

Algorithm 5.1 GAMP

Require: $\mathbf{y}$, $\mathbf{A}$, $\mathbf{S} = \mathbf{A} \cdot \mathbf{A}$, $f_{\mathbf{x}}(\mathbf{x})$, $f_{\mathbf{z}}(\mathbf{z})$

- 1: Initialize: $t = 0$, $\hat{\mathbf{x}}^t$, $\boldsymbol{\tau}_x^t$, $\mathbf{s}^{t-1} = \mathbf{0}$
 - 2: **repeat**
 - 3: [Output node update]
 - 4: $\boldsymbol{\tau}_p^t = \mathbf{S} \boldsymbol{\tau}_x^t$
 - 5: $\mathbf{p}^t = \mathbf{A} \hat{\mathbf{x}}^t - \mathbf{s}^{t-1} \cdot \boldsymbol{\tau}_p^t$
 - 6: $\hat{\mathbf{z}}^t = \mathbb{E}(\mathbf{z} | \mathbf{p}^t, \boldsymbol{\tau}_p^t)$
 - 7: $\boldsymbol{\tau}_z^t = \text{Var}(\mathbf{z} | \mathbf{p}^t, \boldsymbol{\tau}_p^t)$
 - 8: $\mathbf{s}^t = (\hat{\mathbf{z}}^t - \mathbf{p}^t) \cdot \boldsymbol{\tau}_p^t$
 - 9: $\boldsymbol{\tau}_s^t = (\mathbf{1} - \boldsymbol{\tau}_z^t / \boldsymbol{\tau}_p^t) \cdot \boldsymbol{\tau}_p^t$
 - 10: [Input node update]
 - 11: $\boldsymbol{\tau}_r^t = \mathbf{1} / (\mathbf{S}^T \boldsymbol{\tau}_s^t)$
 - 12: $\mathbf{r}^t = \hat{\mathbf{x}}^t + \boldsymbol{\tau}_r^t \cdot \mathbf{A}^T \mathbf{s}^t$
 - 13: $\hat{\mathbf{x}}^{t+1} = \mathbb{E}(\mathbf{x} | \mathbf{r}^t, \boldsymbol{\tau}_r^t)$
 - 14: $\boldsymbol{\tau}_x^{t+1} = \text{Var}(\mathbf{x} | \mathbf{r}^t, \boldsymbol{\tau}_r^t)$
 - 15: **until** Convergence
-

following constrained minimization of a LSL of the BFE (see also [18] and references therein):

$$\begin{aligned}
 & \min_{q_{\mathbf{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p} J_{LSL-BFE}(q_{\mathbf{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p) \\
 & s.t. \quad \mathbb{E}(\mathbf{z} | q_{\mathbf{z}}) = \mathbf{A} \mathbb{E}(\mathbf{x} | q_{\mathbf{x}}) \\
 & \quad \boldsymbol{\tau}_p = \mathbf{S} \text{Var}(\mathbf{x} | q_{\mathbf{x}}),
 \end{aligned} \tag{5.4}$$

where the LSL BFE is given by

$$\begin{aligned}
 J_{LBFE}(q_{\mathbf{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p) &= D(q_{\mathbf{x}} || e^{-f_{\mathbf{x}}}) + D(q_{\mathbf{z}} || e^{-f_{\mathbf{z}}}) \\
 &+ H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p), \text{ with } H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p) = \frac{1}{2} \sum_{k=1}^M \left[\frac{\text{Var}(z_k | q_{z_k})}{\tau_{p_k}} + \ln(2\pi\tau_{p_k}) \right]
 \end{aligned} \tag{5.5}$$

and where $D(q || p) = \mathbb{E}_q(\ln(\frac{q}{p}))$ is the KLD and $H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p)$ is a sum of a KLD and an entropy of Gaussians with identical means but different variances. The LSL BFE optimization problem (5.5) can be reformulated with the following augmented Lagrangian

$$\begin{aligned}
 & \min_{q_{\mathbf{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \mathbf{u}} \max_{\mathbf{s}, \boldsymbol{\tau}_s} L(q_{\mathbf{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \mathbf{u}, \mathbf{s}, \boldsymbol{\tau}_s) \text{ with} \\
 & L = D(q_{\mathbf{x}} || e^{-f_{\mathbf{x}}}) + D(q_{\mathbf{z}} || e^{-f_{\mathbf{z}}}) + H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p) \\
 & + \mathbf{s}^T (\mathbb{E}(\mathbf{z} | q_{\mathbf{z}}) - \mathbf{A} \mathbb{E}(\mathbf{x} | q_{\mathbf{x}})) - \frac{1}{2} \boldsymbol{\tau}_s^T (\boldsymbol{\tau}_p - \mathbf{S} \text{Var}(\mathbf{x} | q_{\mathbf{x}})) \\
 & + \frac{1}{2} \|\mathbb{E}(\mathbf{x} | q_{\mathbf{x}}) - \mathbf{u}\|_{\boldsymbol{\tau}_r}^2 + \frac{1}{2} \|\mathbb{E}(\mathbf{z} | q_{\mathbf{z}}) - \mathbf{A} \mathbf{u}\|_{\boldsymbol{\tau}_p}^2,
 \end{aligned} \tag{5.6}$$

where $\mathbf{s}$, $\boldsymbol{\tau}_s$ are Lagrange multipliers, and $\boldsymbol{\tau}_r = \mathbf{1} / (\mathbf{S}^T \boldsymbol{\tau}_s)$ is just a short-hand notation for a quantity that depends on $\boldsymbol{\tau}_s$. We also use the notations: $\|\mathbf{u}\|_{\boldsymbol{\tau}}^2 = \sum_i u_i^2 / \tau_i$, element-wise multiplication as in $\mathbf{s} \cdot \boldsymbol{\tau}$ and element-wise division as in $\mathbf{1} / \boldsymbol{\tau}$, and $\mathbf{1}$ is a vector of ones. In [44], [45], a careful updating schedule was considered with partial optimization steps on subsets of primal and dual variables. However, that approach is not guaranteed to converge in general. In [46] we continued to consider an alternating optimization approach in which the schedule is less critical and some of the optimizations are reduced to gradient updates. The resulting algorithm can be considered an extended and generalized version of the ADMM algorithm (extended: there are more than two primal variable

groups, generalized: the quadratic augmentation term does not exactly correspond to the linear (mean) constraint). However, there is an alternative point of view, based on [42], where a double mean constraint was introduced leading to the ADMM-GAMP augmented Lagrangian

$$\begin{aligned} \min_{q_x, q_z, \tau_p, \mathbf{u}} \max_{\mathbf{q}, \mathbf{s}, \tau_s} L_A(q_x, q_z, \tau_p, \mathbf{u}, \mathbf{q}, \mathbf{s}, \tau_s) \text{ with} \\ L_A = D(q_x || e^{-f_x}) - \frac{1}{2} \tau_s^T (\tau_p - \mathbf{S} \text{Var}(\mathbf{x}|q_x)) + D(q_z || e^{-f_z}) \\ + H_G(q_z, \tau_p) + \mathbf{q}^T (\mathbb{E}(\mathbf{x}|q_x) - \mathbf{u}) + \mathbf{s}^T (\mathbb{E}(\mathbf{z}|q_z) - \mathbf{A} \mathbf{u}) \\ + \frac{1}{2} \|\mathbb{E}(\mathbf{x}|q_x) - \mathbf{u}\|_{\tau_r}^2 + \frac{1}{2} \|\mathbb{E}(\mathbf{z}|q_z) - \mathbf{A} \mathbf{u}\|_{\tau_p}^2, \end{aligned} \quad (5.7)$$

For ADMM, the first two terms are the cost function for q_x , the next two terms constitute the cost function for q_z . The two groups of primal variables are $\{q_x, q_z\}$ and $\mathbf{u}$ (and the optimization of L_A is decoupled between q_x, q_z). The two linear constraints together constitute a single extended set of linear constraints with extended Lagrange multiplier $[\mathbf{q}^T \mathbf{s}^T]^T$. The appropriately weighted quadratic augmentation terms correspond exactly to the set of linear constraints. The optimization in [42] is organized with the usual ADMM algorithm alternating between minimizations over the two groups of primal variables, followed by the ADMM specific Lagrange multiplier update. The optimization over the remaining variable τ_p, τ_s is then performed in an outer loop. We show here (by the variance subsystem convergence analysis) that this organization in two levels is not necessary. Furthermore, there is a redundancy between the linear and quadratic constraint terms in (5.7). Indeed, if we impose the constrained Lagrange multiplier structure $\mathbf{q}^T = -\mathbf{s}^T \mathbf{A}$, then we obtain the proposed L in (5.6). This is constrained enough since the Lagrange multiplier $\mathbf{s}$ will lead to $\mathbb{E}(\mathbf{z}|q_z) = \mathbf{A} \mathbb{E}(\mathbf{x}|q_x)$, in which case the quadratic augmentation terms are minimized by $\mathbf{u} = \mathbb{E}(\mathbf{x}|q_x)$ and disappear. However, constraining $\mathbf{q}^T = -\mathbf{s}^T \mathbf{A}$ leads to a deviation from the strict ADMM structure and requires separate convergence analysis, which we provide here.

At iteration t we propose the following updating sequence

$$\{\mathbf{u}^t\} = \arg \min_{\mathbf{u}} L(q_x^{t-1}, q_z^{t-1}, \tau_p^{t-1}, \mathbf{u}, \mathbf{s}^{t-1}, \tau_s^{t-1}) \quad (5.8)$$

$$\{q_x^t\} = \arg \min_{q_x} L(q_x, q_z^{t-1}, \tau_p^{t-1}, \mathbf{u}^t, \mathbf{s}^{t-1}, \tau_s^{t-1}) \quad (5.9)$$

$$\{q_z^t\} = \arg \min_{q_z} L(q_x^t, q_z, \tau_p^t, \mathbf{u}^t, \mathbf{s}^{t-1}, \tau_s^{t-1}) \quad (5.10)$$

$$\{\mathbf{s}^t\} = \arg \max_{\mathbf{s}} L(q_x^t, q_z^t, \tau_p^t, \mathbf{u}^t, \mathbf{s}, \tau_s^{t-1}) \quad (5.11)$$

$$\{\tau_p^t, \tau_s^t\} = \arg \min_{\tau_p} \max_{\tau_s} L(q_x^t, q_z^t, \tau_p, \mathbf{u}^t, \mathbf{s}^t, \tau_s) \quad (5.12)$$

The result appears in Algorithm 5.2.

5.3.1 Update of $\mathbf{u}$

To update $\mathbf{u}$, we use a gradient descent method with line search to optimize the step-size. From (5.6), (5.8), we get

$$\begin{aligned} L(q_x^{t-1}, q_z^{t-1}, \tau_p^{t-1}, \mathbf{u}, \mathbf{s}^{t-1}, \tau_s^{t-1}) \\ = \frac{1}{2} \|\hat{\mathbf{x}}^{t-1} - \mathbf{u}\|_{\tau_r^{t-1}}^2 + \frac{1}{2} \|\hat{\mathbf{z}}^{t-1} - \mathbf{A} \mathbf{u}\|_{\tau_p^{t-1}}^2 + \text{const.} \end{aligned} \quad (5.13)$$

where *const.* denotes constants w.r.t. $\mathbf{u}$. The minimizing update can be obtained as

$$\mathbf{u}^t = \mathbf{u}^{t-1} - \eta^t \mathbf{g}^t \quad (5.14)$$

with gradient $\mathbf{g}^t = \mathbf{g}^t(\mathbf{u}^{t-1})$ where

$$\begin{aligned} \mathbf{g}^t(\mathbf{u}) &= \nabla_{\mathbf{u}} L(q_x^{t-1}, q_z^{t-1}, \tau_p^{t-1}, \mathbf{u}, \mathbf{s}^{t-1}, \tau_s^{t-1}) \\ &= -\mathbf{A}^T ((\hat{\mathbf{z}}^{t-1} - \mathbf{A} \mathbf{u}) ./ \tau_p^{t-1}) - (\hat{\mathbf{x}}^{t-1} - \mathbf{u}) ./ \tau_r^{t-1} \\ &= \mathbf{g}^t(0) + \mathcal{H}^t \mathbf{u}, \quad \mathcal{H}^t = \mathbf{D}(1./\tau_r^{t-1}) + \mathbf{A}^T \mathbf{D}(1./\tau_p^{t-1}) \mathbf{A} \end{aligned} \quad (5.15)$$

where $\mathbf{D}(\boldsymbol{\tau})$ denotes a diagonal matrix with diagonal elements $\boldsymbol{\tau}$. The step-size η^t gets optimized for maximum descent :

$$\begin{aligned} & \frac{\partial L(q_{\mathbf{x}}^{t-1}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \mathbf{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1})}{\partial \eta^t} = 0 \\ \Rightarrow \quad \eta^t &= \|\mathbf{g}^t\|^2 / \mathbf{g}^{tT} \mathcal{H}^t \mathbf{g}^t. \end{aligned} \quad (5.16)$$

5.3.2 Update of $q_{\mathbf{x}}$

For the update of $q_{\mathbf{x}}$, consider the relevant terms in the augmented Lagrangian (and remember that $\mathbf{1}./\boldsymbol{\tau}_r^{t-1} = \mathbf{S}^T \boldsymbol{\tau}_s^{t-1}$)

$$\begin{aligned} & L(q_{\mathbf{x}}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \mathbf{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \\ &= D(q_{\mathbf{x}} || e^{-f_{\mathbf{x}}}) - \mathbf{s}^{t-1T} \mathbf{A} \mathbb{E}(\mathbf{x} | q_{\mathbf{x}}) + \frac{1}{2} \boldsymbol{\tau}_s^{t-1T} \mathbf{S} \text{Var}(\mathbf{x} | q_{\mathbf{x}}) + \frac{1}{2} \|\mathbb{E}(\mathbf{x} | q_{\mathbf{x}}) - \mathbf{u}^t\|_{\boldsymbol{\tau}_r^{t-1}}^2 + \text{const.} \\ &= D(q_{\mathbf{x}} || e^{-f_{\mathbf{x}}}) + \frac{1}{2} (\mathbf{1}./\boldsymbol{\tau}_r^{t-1})^T \mathbb{E}(\mathbf{x} \cdot \mathbf{x} | q_{\mathbf{x}}) - \mathbf{s}^{t-1T} \mathbf{A} \mathbb{E}(\mathbf{x} | q_{\mathbf{x}}) - (\mathbf{u}^t ./ \boldsymbol{\tau}_r^{t-1})^T \mathbb{E}(\mathbf{x} | q_{\mathbf{x}}) + \text{const.} \\ &= D(q_{\mathbf{x}} || e^{-f_{\mathbf{x}}}) + \frac{1}{2} (\mathbf{1}./\boldsymbol{\tau}_r^{t-1})^T \mathbb{E}(\mathbf{x} \cdot \mathbf{x} | q_{\mathbf{x}}) - (\mathbf{u}^t + \boldsymbol{\tau}_r^{t-1} \cdot \mathbf{A}^T \mathbf{s}^{t-1})^T (\mathbb{E}(\mathbf{x} | q_{\mathbf{x}}) ./ \boldsymbol{\tau}_r^{t-1}) + \text{const.} \\ &= D(q_{\mathbf{x}} || e^{-f_{\mathbf{x}}}) + \frac{1}{2} \mathbb{E}(\|\mathbf{x} - \mathbf{r}^t\|_{\boldsymbol{\tau}_r^t}^2 | q_{\mathbf{x}}) + \text{const.} \end{aligned} \quad (5.17)$$

where const. denotes constants w.r.t. $\mathbf{x}$, and $\mathbf{r}^t = \mathbf{u}^t + \boldsymbol{\tau}_r^{t-1} \cdot \mathbf{A}^T \mathbf{s}^{t-1}$. The Lagrangian in (5.17) is separable. We get per component

$$\begin{aligned} \min_{q_{x_i}} D(q_{x_i} || g_{x_i}^t / Z_{x_i}^t) &\Rightarrow q_{x_i}^t = g_{x_i}^t / Z_{x_i}^t, \quad Z_{x_i}^t = \int g_{x_i}^t(x_i) dx_i, \\ -\ln g_{x_i}^t(x_i) &= f_{x_i}(x_i) + \frac{1}{2\tau_{r_i}^t} [(x_i - r_i^t)^2 - r_i^{t2}]. \end{aligned} \quad (5.18)$$

The partition function $Z_{x_i}^t$ acts as cumulant generating function:

$$\begin{aligned} \tau_{r_i}^t \frac{\partial \ln Z_{x_i}^t}{\partial r_i^t} &= \mathbb{E}(x_i | q_{x_i}^t) = \mathbb{E}(x_i | r_i^t, \tau_{r_i}^t) = \hat{x}_i^t \\ (\tau_{r_i}^t)^2 \frac{\partial^2 \ln Z_{x_i}^t}{\partial r_i^{t2}} &= \text{Var}(x_i | r_i^t, \tau_{r_i}^t) = \tau_{x_i}^t. \end{aligned} \quad (5.19)$$

In the Gaussian prior case, we get a Gaussian posterior $q_{\mathbf{x}}^t$ with

$$\mathbf{1}./\boldsymbol{\tau}_x^t = \mathbf{1}./\boldsymbol{\tau}_r^{t-1} + \mathbf{1}./\boldsymbol{\sigma}_x^2, \quad \hat{\mathbf{x}}^t = \boldsymbol{\tau}_x^t \cdot (\mathbf{r}^t ./ \boldsymbol{\tau}_r^{t-1}). \quad (5.20)$$

5.3.3 Update of $\{q_{\mathbf{z}}\}$

The relevant terms in the augmented Lagrangian are

$$\begin{aligned} & L(q_{\mathbf{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p^{t-1}, \mathbf{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \\ &= D(q_{\mathbf{z}} || e^{-f_{\mathbf{z}}}) + \frac{1}{2} \text{Var}(\mathbf{z} | q_{\mathbf{z}}) ./ \boldsymbol{\tau}_p^{t-1} \\ &+ \mathbf{s}^{t-1T} \mathbb{E}(\mathbf{z} | q_{\mathbf{z}}) + \frac{1}{2} \|\mathbb{E}(\mathbf{z} | q_{\mathbf{z}}) - \mathbf{A} \mathbf{u}^t\|_{\boldsymbol{\tau}_p^{t-1}}^2 + \text{const.} \\ &= D(q_{\mathbf{z}} || e^{-f_{\mathbf{z}}}) + \frac{1}{2} \mathbb{E}(\mathbf{z}^T \mathbf{z} | q_{\mathbf{z}}) ./ \boldsymbol{\tau}_p^{t-1} \\ &- (\mathbb{E}(\mathbf{z} | q_{\mathbf{z}}))^T ((\mathbf{A} \mathbf{u}^t) ./ \boldsymbol{\tau}_p^{t-1} - \mathbf{s}^{t-1}) + \text{const.} \\ &= D(q_{\mathbf{z}} || e^{-f_{\mathbf{z}}}) + \frac{1}{2} \mathbb{E}(\|\mathbf{z} - \mathbf{p}^t\|_{\boldsymbol{\tau}_p^{t-1}}^2 | q_{\mathbf{z}}) + \text{const.} \end{aligned} \quad (5.21)$$

where *const.* denotes constants w.r.t. $\mathbf{z}$ and $\mathbf{p}^t = \mathbf{A}\mathbf{u}^t - \mathbf{s}^{t-1} \cdot \boldsymbol{\tau}_p^{t-1}$. The Lagrangian in (5.21) is again separable. We get per component

$$\begin{aligned} \min_{q_{z_k}} D(q_{z_k} || g_{z_k}^t / Z_{z_k}^t) &\Rightarrow q_{z_k}^t = g_{z_k}^t / Z_{z_k}^t \\ Z_{z_k}^t &= \int g_{z_k}^t(z_k) dz_k, \quad -\ln g_{z_k}^t(z_k) = \\ f_{z_k}(z_k) &+ \frac{1}{2\tau_{p_k}^{t-1}} [(z_k - p_k^t)^2 - (p_k^t)^2]. \end{aligned} \quad (5.22)$$

The partition function $Z_{z_k}^t$ acts again as cumulant generating function:

$$\begin{aligned} -\frac{\partial \ln Z_{z_k}^t}{\partial s_k^{t-1}} &= \mathbb{E}(z_k | q_{z_k}^t) = \mathbb{E}(z_k | p_k^t, \tau_{p_k}^{t-1}) = \hat{z}_k^t \\ \frac{\partial^2 \ln Z_{z_k}^t}{\partial s_k^{t-1 2}} &= \text{Var}(z_k | p_k^t, \tau_{p_k}^{t-1}) = \tau_{z_k}^t \\ -\frac{\partial^3 \ln Z_{z_k}^t}{\partial s_k^{t-1 3}} &= \mathbb{E}((z_k - \mathbb{E}z_k)^3 | q_{z_k}^t). \end{aligned} \quad (5.23)$$

The case of Gaussian noise leads again to a Gaussian posterior $q_{\mathbf{z}}$ with

$$\mathbf{1} \cdot \boldsymbol{\tau}_z^t = \mathbf{1} \cdot \boldsymbol{\tau}_p^{t-1} + \mathbf{1} \cdot \boldsymbol{\sigma}_v^2, \quad \hat{\mathbf{z}}^t = \boldsymbol{\tau}_z^t \cdot (\mathbf{y} \cdot \boldsymbol{\sigma}_v^2 + \mathbf{p}^t \cdot \boldsymbol{\tau}_p^{t-1}). \quad (5.24)$$

5.3.4 Update of $\{\mathbf{s}\}$ (ADMM style)

Although the quadratic augmentation terms in the Lagrangian do not correspond exactly to a weighted quadratic version of the linear mean constraint, due to the introduction of the auxiliary variable $\mathbf{u}$ which streamlines the derivation of the updates of $q_{\mathbf{x}}$ and $q_{\mathbf{z}}$, nevertheless an ADMM style update of the mean constraint Lagrange multiplier $\mathbf{s}$ is possible. Indeed, the terms in (5.21) that contains $\mathbf{s}$ or $\hat{\mathbf{z}}$ are

$$\hat{\mathbf{z}}^T ((\frac{1}{2}\hat{\mathbf{z}} - \mathbf{p}^t) \cdot \boldsymbol{\tau}_p^{t-1}) = \hat{\mathbf{z}}^T (\mathbf{s}^{t-1} + (\frac{1}{2}\hat{\mathbf{z}} - \mathbf{A}\mathbf{u}^t) \cdot \boldsymbol{\tau}_p^{t-1}) \quad (5.25)$$

Taking the gradient w.r.t. $\hat{\mathbf{z}}$ (as part of the optimization over $q_{\mathbf{z}}$) leads to the RHS of

$$\mathbf{s}^t = \mathbf{s}^{t-1} + (\hat{\mathbf{z}}^t - \mathbf{A}\mathbf{u}^t) \cdot \boldsymbol{\tau}_p^{t-1}. \quad (5.26)$$

Hence, if we use this update for $\mathbf{s}$, then (5.25) reduces to $\hat{\mathbf{z}}^T \mathbf{s}^t$, as if the quadratic augmentation terms have disappeared! This is the main characteristic of the Lagrange multiplier update in ADMM, which corresponds to a gradient ascent with a particular choice of (diagonal) step-size.

5.3.5 Update of $\{\boldsymbol{\tau}_p, \boldsymbol{\tau}_s\}$

In [44], [45], the carefully chosen updating schedule made the quadratic augmentation terms inactive when updating $\{\boldsymbol{\tau}_p, \boldsymbol{\tau}_s\}$. Here these terms only become inactive at convergence. Nevertheless, these terms only play an active role for the means and not for the variances. Hence, we shall ignore them here. Thus, the terms of interest in (5.6) for (5.12) are

$$\begin{aligned} &L(q_{\mathbf{x}}^t, q_{\mathbf{z}}^t, \boldsymbol{\tau}_p, \mathbf{u}^t, \mathbf{s}^t, \boldsymbol{\tau}_s) \\ &= H_G(q_{\mathbf{z}}^t, \boldsymbol{\tau}_p) - \frac{1}{2} \boldsymbol{\tau}_s^T (\boldsymbol{\tau}_p - \mathbf{S} \boldsymbol{\tau}_x^t) + \text{const.} = \text{const.} + \\ &\frac{1}{2} \sum_{k=1}^M \left[\frac{\tau_{z_k}^t}{\tau_{p_k}} + \ln(2\pi \tau_{p_k}) \right] - \frac{1}{2} \sum_{k=1}^M \tau_{s_k} (\tau_{p_k} - \mathbf{S}_{k,:} \boldsymbol{\tau}_x^t) \end{aligned} \quad (5.27)$$

where *const.* denotes constants w.r.t. $\{\boldsymbol{\tau}_p, \boldsymbol{\tau}_s\}$. Deriving w.r.t. $\{\boldsymbol{\tau}_p, \boldsymbol{\tau}_s\}$ yields the feasibility conditions

$$\frac{\partial L}{\partial \tau_{s_k}} = 0 \Rightarrow \tau_{p_k}^t = \mathbf{S}_{k,:} \tau_x^t \quad (5.28)$$

$$\frac{\partial L}{\partial \tau_{p_k}} = 0 \Rightarrow \tau_{s_k}^t = \frac{1}{\tau_{p_k}^t} \left(1 - \frac{\tau_{z_k}^t}{\tau_{p_k}^t}\right). \quad (5.29)$$

which we run as a fixed-point sub-algorithm. The position of these updates in the updating schedule is less important. Nevertheless we shall update $\{\tau_p, \tau_s\}$ as soon as the quantities on which they depend have been updated.

Algorithm 5.2 AMBGAMP

Require: $\mathbf{y}, \mathbf{A}, \mathbf{S} = \mathbf{A} \cdot \mathbf{A}, f_x(\mathbf{x}), f_z(\mathbf{z})$

- 1: Initialize: $t = 0, \mathbf{u}^0 = \mathbf{0}, \hat{\mathbf{x}}^0 = \mathbf{0}, \hat{\mathbf{z}}^0 = \mathbf{0}, \mathbf{s}^0 = \mathbf{0}, \tau_r^0 = \mathbf{1}, \tau_p^0 = \mathbf{1}$
 - 2: **repeat** ($t=1,2,\dots$)
 - 3: $\mathbf{u}^t = \mathbf{u}^{t-1} - \eta^t \mathbf{g}^t$, with $\mathbf{g}^t, \eta^t$ from (5.15), (5.16)
 - 4: [Input node update]
 - 5: $\mathbf{r}^t = \mathbf{u}^t + \tau_r^{t-1} \cdot (\mathbf{A}^T \mathbf{s}^{t-1})$
 - 6: $\hat{\mathbf{x}}^t = \mathbb{E}(\mathbf{x} | \mathbf{r}^t, \tau_r^{t-1})$, Gaussian p_x : $\mathbf{1} \cdot / \tau_x^t = \mathbf{1} \cdot / \tau_r^{t-1} + \mathbf{1} \cdot / \sigma_x^2$
 - 7: $\tau_x^t = \text{Var}(\mathbf{x} | \mathbf{r}^t, \tau_r^{t-1})$, Gaussian p_x : $\hat{\mathbf{x}}^t = \tau_x^t \cdot (\mathbf{r}^t \cdot / \tau_r^{t-1})$
 - 8: $\tau_p^t = \mathbf{S} \tau_x^t$
 - 9: [Output node update]
 - 10: $\mathbf{p}^t = \mathbf{A} \mathbf{u}^t - \mathbf{s}^{t-1} \cdot \tau_p^t$
 - 11: $\hat{\mathbf{z}}^t = \mathbb{E}(\mathbf{z} | \mathbf{p}^t, \tau_p^t)$, Gaussian $p_{y|z}$: $\mathbf{1} \cdot / \tau_z^t = \mathbf{1} \cdot / \tau_p^t + \mathbf{1} \cdot / \sigma_v^2$
 - 12: $\tau_z^t = \text{Var}(\mathbf{z} | \mathbf{p}^t, \tau_p^t)$, Gaussian $p_{y|z}$: $\hat{\mathbf{z}}^t = \tau_z^t \cdot (\mathbf{y} \cdot / \sigma_v^2 + \mathbf{p}^t \cdot / \tau_p^t)$
 - 13: $\mathbf{s}^t = \mathbf{s}^{t-1} + (\hat{\mathbf{z}}^t - \mathbf{A} \mathbf{u}^t) \cdot / \tau_p^t$
 - 14: $\tau_s^t = (\mathbf{1} - \tau_z^t \cdot / \tau_p^t) \cdot / \tau_p^t$, Gaussian $p_{y|z}$: $\tau_s^t = \mathbf{1} \cdot / (\sigma_v^2 + \tau_p^t)$
 - 15: $\tau_r^t = \mathbf{1} \cdot / (\mathbf{S}^T \tau_s^t)$
 - 16: **until** Convergence
-

5.4 Convergence to LMMSE

In the case of Gaussian $p_x, p_{y|z}$, the cost function is quadratic in $\mathbf{x}$ etc., and we check convergence to the LMMSE estimate. At convergence we have

$$\begin{aligned}
\mathbf{s}^t &= \mathbf{s}^{t-1} \Rightarrow \hat{\mathbf{z}} = \mathbf{A} \mathbf{u} \\
&\Rightarrow \mathbf{u} = \hat{\mathbf{x}} \text{ from } \mathbf{g}(\mathbf{u}) = \mathbf{0} = \mathbf{g}(\mathbf{0}) + \mathcal{H} \mathbf{u} \\
\hat{\mathbf{z}} &= \tau_z \cdot (\mathbf{y} \cdot / \sigma_v^2 + (\mathbf{A} \mathbf{u}) \cdot / \tau_p - \mathbf{s}) \\
&\Rightarrow \mathbf{s} = \mathbf{y} \cdot / \sigma_v^2 + (\mathbf{A} \hat{\mathbf{x}}) \cdot / \tau_p - (\mathbf{A} \hat{\mathbf{x}}) \cdot / \tau_z = (\mathbf{y} - \mathbf{A} \hat{\mathbf{x}}) \cdot / \sigma_v^2 \\
\hat{\mathbf{x}} &= (\tau_x \cdot / \tau_r) \cdot (\mathbf{u} + \tau_r \cdot (\mathbf{A}^T \mathbf{s})) \\
&\Rightarrow (\mathbf{1} - \tau_x \cdot / \tau_r) \hat{\mathbf{x}} = \tau_x \cdot \hat{\mathbf{x}} \cdot / \sigma_x^2 = \tau_x \cdot (\mathbf{A}^T \mathbf{s}) \text{ or} \\
\hat{\mathbf{x}} &= [\mathbf{A}^T \mathbf{D}^{-1}(\sigma_v^2) \mathbf{A} + \mathbf{D}^{-1}(\sigma_x^2)]^{-1} \mathbf{A}^T \mathbf{D}^{-1}(\sigma_v^2) \mathbf{y}.
\end{aligned} \quad (5.30)$$

Note that at convergence $\hat{\mathbf{x}}$ does not depend on the various variance estimates that the algorithm produces. One can also get the following convergence values

$$\mathbf{s} = \mathbf{R}_{yy}^{-1} \mathbf{y}, \hat{\mathbf{x}} = \mathbf{D}(\sigma_x^2) \mathbf{A}^T \mathbf{s}, \mathbf{r} = \mathbf{D}(\sigma_x^2 + \tau_r) \mathbf{A}^T \mathbf{s} \quad (5.31)$$

where $\mathbf{r}$ corresponds to the componentwise conditionally unbiased MMSE estimate of $\mathbf{x}$ [24], [47] if τ_r converges to its correct value.

Below we shall analyze the convergence of the proposed AMBGAMP algorithm. Note that the updates of $q_{\mathbf{x}}, q_{\mathbf{z}}$ in (5.18), (5.22) imply that these approximate posteriors inherit the higher order cumulants of their respective priors (cf. Edgeworth expansions around a Gaussian). Only the means and variances are affected by the iterative algorithm. In the Gaussian case, the mean subsystem depends on the variances, but the variance subsystem runs independently. Hence their convergence can be analyzed separately. In the non-Gaussian case, their coupling may need to be reconsidered though.

5.5 Convergence of the Variance Subsystem

In the Gaussian priors case, the updates of the variances can be checked to result in the following variance subsystem

$$\begin{aligned} \mathbf{1}./\boldsymbol{\tau}_x^t &= \mathbf{1}./\boldsymbol{\sigma}_x^2 + \mathbf{S}^T \boldsymbol{\tau}_s^{t-1}, \\ \mathbf{1}./\boldsymbol{\tau}_s^t &= \boldsymbol{\sigma}_v^2 + \mathbf{S} \boldsymbol{\tau}_x^t. \end{aligned} \quad (5.32)$$

To analyze convergence, we investigate the contractiveness of the mappings via their Jacobians

$$\frac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_s^{t-1T}} = -\mathbf{D}_x^{t,2} \mathbf{S}^T, \quad \frac{\partial \boldsymbol{\tau}_s^t}{\partial \boldsymbol{\tau}_x^{tT}} = -\mathbf{D}_s^{t,2} \mathbf{S} \quad (5.33)$$

where we introduced the notation $\mathbf{D}_x^t = \mathbf{D}(\boldsymbol{\tau}_x^t)$ etc., $\mathbf{D}_x^{t,2} = (\mathbf{D}_x^t)^2$ etc. By the chain rule, we get

$$\begin{aligned} \frac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_s^{t-1T}} &= \frac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_s^{t-1T}} \frac{\partial \boldsymbol{\tau}_s^{t-1}}{\partial \boldsymbol{\tau}_x^{t-1T}} \\ \frac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_x^{1T}} &= \mathbf{D}_x^{t,2} \mathbf{S}^T \mathbf{D}_s^{t-1,2} \mathbf{S} \mathbf{D}_x^{t-1,2} \mathbf{S}^T \mathbf{D}_s^{t-2,2} \mathbf{S} \dots \end{aligned} \quad (5.34)$$

Note that the cascade of Jacobians involves a cascade of the following matrices

$$\mathbf{B}^t = \mathbf{B}_x^t \mathbf{B}_s^{t-1} = (\mathbf{D}_x^t \mathbf{S}^T \mathbf{D}_s^{t-1}) (\mathbf{D}_s^{t-1} \mathbf{S} \mathbf{D}_x^{t-1}) \quad (5.35)$$

We can investigate the contractivity of $\mathbf{B}^t$ using any norm since all valid norms are commensurate. A judicious choice here is the infinity norm

$$\|\mathbf{B}^t\|_\infty = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{B}^t \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{\|\mathbf{x}\|_\infty=1} \|\mathbf{B}^t \mathbf{x}\|_\infty = \|\mathbf{B}^t \mathbf{1}\|_\infty \quad (5.36)$$

where the last identity follows from the non-negativity of the elements of $\mathbf{B}$ (and $\mathbf{B}_x$ or $\mathbf{B}_s$) which implies that $\mathbf{B}^t \mathbf{x} \leq \mathbf{B}^t \mathbf{1}$ for any $\mathbf{x}$ with $\|\mathbf{x}\|_\infty = 1$ (where the relation $\mathbf{x} \leq \mathbf{y}$ indicates that $\mathbf{x}$ is element-wise not larger than $\mathbf{y}$). Now we have

$$\begin{aligned} \mathbf{B}^t \mathbf{1} &= \mathbf{B}_x^t \mathbf{B}_s^{t-1} \mathbf{1} \leq \|\mathbf{B}_s^{t-1}\|_\infty \mathbf{B}_x^t \mathbf{1} \leq \|\mathbf{B}_x^t\|_\infty \|\mathbf{B}_s^{t-1}\|_\infty \mathbf{1} \\ \Rightarrow \|\mathbf{B}^t\|_\infty &\leq \|\mathbf{B}_x^t\|_\infty \|\mathbf{B}_s^{t-1}\|_\infty < 1 \end{aligned} \quad (5.37)$$

where we assume that at least one of $\|\mathbf{B}_x^t\|_\infty \leq 1$, $\|\mathbf{B}_s^{t-1}\|_\infty \leq 1$ is strictly smaller than one. So, (5.37) implies convergence of the variance subsystem. The statements in (5.37) hold in the general non-Gaussian case. In the Gaussian case, we get from (5.35), (5.32)

$$\begin{aligned} \|\mathbf{B}_s^{t-1}\|_\infty &= \|\mathbf{B}_s^{t-1} \mathbf{1}\|_\infty = \|\mathbf{D}_s^{t-1} \mathbf{S} \mathbf{D}_x^{t-1} \mathbf{1}\|_\infty \\ &= \|\mathbf{D}_s^{t-1} \mathbf{S} \boldsymbol{\tau}_x^{t-1}\|_\infty = \max_k \frac{\mathbf{S}_{k,:} \boldsymbol{\tau}_x^{t-1}}{\sigma_{v,k}^2 + \mathbf{S}_{k,:} \boldsymbol{\tau}_x^{t-1}} \leq 1, \\ \|\mathbf{B}_x^t\|_\infty &= \|\mathbf{B}_x^t \mathbf{1}\|_\infty = \|\mathbf{D}_x^t \mathbf{S}^T \mathbf{D}_s^{t-1} \mathbf{1}\|_\infty \\ &= \|\mathbf{D}_x^t \mathbf{S}^T \boldsymbol{\tau}_s^{t-1}\|_\infty = \max_i \frac{\mathbf{S}_{:,i} \boldsymbol{\tau}_s^{t-1}}{1/\sigma_{x,i}^2 + \mathbf{S}_{:,i} \boldsymbol{\tau}_s^{t-1}} \leq 1. \end{aligned} \quad (5.38)$$

5.6 Large System Analysis with n.i.i.d. $\mathbf{A}$

To show that at convergence, the variance subsystem $\boldsymbol{\tau}_x$ converges in the large system limit to the optimal MSE in the Gaussian case. We use the following result from [48], [49]:

Theorem 1. *Let $\mathbf{Q}_N, \mathbf{D}_N \in \mathbb{R}^{N \times N}$ be deterministic symmetric matrices and $\mathbf{Y}_N = \mathbf{X}_N \mathbf{D} \mathbf{X}_N^H = \sum_{i=1}^M d_i \mathbf{x}_i \mathbf{x}_i^H$, with diagonal $\mathbf{D}$ and $\mathbf{X}_N$ containing M independent columns $\mathbf{x}_i$ with covariance matrix $\boldsymbol{\Theta}_i$. Also, assume that $\mathbf{Q}_N, \mathbf{D}_N, \boldsymbol{\Theta}_i$ have uniformly bounded spectral norms. Then, as $M, N \rightarrow \infty$ at constant ratio*

$$\frac{1}{N} \text{tr} [\mathbf{Q}_N (\mathbf{Y}_N + \mathbf{D}_N)^{-1}] - \frac{1}{N} \text{tr} [\mathbf{Q}_N \mathbf{T}_N] \xrightarrow{a.s.} 0, \text{ with} \quad (5.39)$$

$$\mathbf{T}_N = \left(\sum_{i=1}^M \frac{d_i \boldsymbol{\Theta}_i}{1 + e_i} + \mathbf{D}_N \right)^{-1}, \text{ where the } e_i \text{ satisfy} \quad (5.40)$$

$$e_k = \text{tr} \left[d_k \boldsymbol{\Theta}_k \left(\sum_{i=1}^M \frac{d_i \boldsymbol{\Theta}_i}{1 + e_i} + \mathbf{D}_N \right)^{-1} \right], k = 1, \dots, M. \quad (5.41)$$

The convergence in (5.39) is the convergence of a scalar to its mean, by LLN. Note the presence of the weights $1 + e_i$ in the denominator of the sum in $\mathbf{T}_N$ in (5.40), which reflect that the expected value of a matrix inverse is not the inverse of its expected value. Note that the $\text{tr} [\boldsymbol{\Theta}_i]$ should be of order 1, which means that the sum in $\mathbf{T}_N$ is implicitly normalized. The e_i satisfy the implicit equations (5.41), and can be obtained as the fixed points of the RHS interpreted as a mapping (with global convergence).

We assume the columns of $\mathbf{A}^T = [\mathbf{a}_1 \dots \mathbf{a}_M]$ to be zero mean and independent with diagonal covariance matrix $\mathbb{E}(\mathbf{a}_i \mathbf{a}_i^T) = \boldsymbol{\Theta}_i$. The optimal MSE in the Gaussian case is given by (with e.g. $\mathbf{D}_{\sigma_x^2} = \mathbf{D}(\sigma_x^2)$)

$$\begin{aligned} \text{MSE} &= \frac{1}{N} \text{tr} \left\{ \left[\mathbf{A}^T \mathbf{D}_{\sigma_v^2}^{-1} \mathbf{A} + \mathbf{D}_{\sigma_x^2}^{-1} \right]^{-1} \right\} \\ &\xrightarrow{a.s.} \frac{1}{N} \text{tr} \left\{ \left[\sum_{i=1}^M \frac{\boldsymbol{\Theta}_i}{\sigma_{v,i}^2 (1 + e_i)} + \mathbf{D}_{\sigma_x^2}^{-1} \right]^{-1} \right\}, \text{ with} \end{aligned} \quad (5.42)$$

$$e_k = \text{tr} \left\{ \frac{1}{\sigma_{v,k}^2} \boldsymbol{\Theta}_k \left[\sum_{i=1}^M \frac{\boldsymbol{\Theta}_i}{\sigma_{v,i}^2 (1 + e_i)} + \mathbf{D}_{\sigma_x^2}^{-1} \right]^{-1} \right\}. \quad (5.43)$$

On the other hand the GAMP variance subsystem converges to (5.32), without iteration indices. With large $\mathbf{A}$, $\mathbf{S} \boldsymbol{\tau}_x$ and $\mathbf{S}^T \boldsymbol{\tau}_s$ converge to their expected values

$$\mathbb{E} [\mathbf{S} \boldsymbol{\tau}_x]_i = \mathbb{E} [\mathbf{A} \mathbf{D}_x \mathbf{A}^T]_{ii} = \text{tr} \{ \boldsymbol{\Theta}_i \mathbf{D}_x \}; \quad (5.44)$$

$$\mathbb{E} \mathbf{D} (\mathbf{S}^T \boldsymbol{\tau}_s) = \mathbb{E} \text{diag} (\mathbf{A}^T \mathbf{D}_s \mathbf{A}) = \sum_{k=1}^M \tau_{s,k} \boldsymbol{\Theta}_k. \quad (5.45)$$

Therefore, the empirical mean of the posterior variance $\boldsymbol{\tau}_x$ becomes

$$\frac{1}{N} \text{tr} \{ \mathbf{D}_x \} = \frac{1}{N} \text{tr} \left\{ \left[\mathbf{D}_{\sigma_x^2}^{-1} + \sum_{k=1}^M \tau_{s,k} \boldsymbol{\Theta}_k \right]^{-1} \right\}. \quad (5.46)$$

From (5.32), (5.44), it follows that

$$\tau_{s,k} = \frac{1}{\sigma_{v,k} + \text{tr}\{\mathbf{\Theta}_k \mathbf{D}_x\}}. \quad (5.47)$$

Define $e'_k = \frac{\text{tr}\{\mathbf{\Theta}_k \mathbf{D}_x\}}{\sigma_{v,k}^2}$ and substituting (5.47) into (5.46), we obtain

$$\begin{aligned} \frac{1}{N} \text{tr}\{\mathbf{D}(\boldsymbol{\tau}_x)\} &= \frac{1}{N} \text{tr} \left\{ \left[\mathbf{D}_{\sigma_x^2}^{-1} + \sum_{i=1}^M \frac{\mathbf{\Theta}_i}{\sigma_{v,i}(1 + e'_i)} \right]^{-1} \right\}; \\ e'_k &= \text{tr} \left\{ \frac{\mathbf{\Theta}_k}{\sigma_{v,k}^2} \left[\mathbf{D}_{\sigma_x^2}^{-1} + \sum_{i=1}^M \frac{\mathbf{\Theta}_i}{\sigma_{v,i}(1 + e'_i)} \right]^{-1} \right\}. \end{aligned} \quad (5.48)$$

This shows that the system of equations defined by (5.48) is the same as the system defined by (5.42), (5.43). Therefore, the empirical mean of $\boldsymbol{\tau}_x$ converges to the optimal MSE in the large system limit. Note that above we have applied the Theorem with $\mathbf{Q}_N = \mathbf{I}_N$ but the same results hold for any $\mathbf{Q}_N$, corresponding to deterministic limits for variably weighted MSEs $\text{tr}\{\mathbf{Q}_N \mathbf{D}_x\}/\text{tr}\{\mathbf{Q}_N\}$.

5.7 Convergence of the Mean Subsystem

We consider here the convergence proof for the case in which the update of $\mathbf{u}$ minimizes its quadratic cost function, i.e. $\mathbf{g}^t(\mathbf{u}^t) = \mathbf{g}^t(\mathbf{0}) + \mathcal{H}^t \mathbf{u}^t = \mathbf{0} \Rightarrow \mathbf{u}^t = -(\mathcal{H}^t)^{-1} \mathbf{g}^t(\mathbf{0})$. We shall investigate the convergence of the mean subsystem once the variance subsystem has converged. Similar to the convergence proof in [42], we will derive the Jacobian of the updating function and prove the convergence by showing that the Jacobian is contractive. We define the short hand notations

$$\begin{aligned} \boldsymbol{\tau} &= [\boldsymbol{\tau}_r^T \quad \boldsymbol{\tau}_p^T]^T, \mathbf{D} = \mathbf{D}(1./\boldsymbol{\tau}), \mathbf{B} = \mathbf{D}^{\frac{1}{2}} [\mathbf{I} \quad \mathbf{A}^T]^T, \\ \mathbf{C} &= \mathbf{D}^{-\frac{1}{2}} [\mathbf{A} \quad -\mathbf{I}]^T, \mathbf{H} = [\mathbf{0}_{M \times N} \quad \mathbf{I}] \mathbf{D}^{\frac{1}{2}}, \\ \mathbf{w}^t &= \mathbf{D}^{\frac{1}{2}} [\hat{\mathbf{x}}^{tT} \quad \hat{\mathbf{z}}^{tT}]^T, \mathbf{P} = \mathbf{B} (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T. \end{aligned} \quad (5.49)$$

Furthermore, we define the update function for $[\hat{\mathbf{x}}^{tT} \quad \hat{\mathbf{z}}^{tT}]^T$ as

$$\begin{bmatrix} \hat{\mathbf{x}}^t \\ \hat{\mathbf{z}}^t \end{bmatrix} = \mathbf{g} \left(\begin{bmatrix} \mathbf{r}^t \\ \mathbf{p}^t \end{bmatrix} \right) = \begin{bmatrix} \mathbf{g}_x(\mathbf{r}^t) \\ \mathbf{g}_z(\mathbf{p}^t) \end{bmatrix} = \begin{bmatrix} \mathbb{E}(\mathbf{x}|\mathbf{r}^t, \boldsymbol{\tau}_r) \\ \mathbb{E}(\mathbf{z}|\mathbf{p}^t, \boldsymbol{\tau}_p) \end{bmatrix}. \quad (5.50)$$

In the following we will show that the system $\boldsymbol{\theta}^t = [\mathbf{w}^{t,T} \quad \mathbf{s}^{t,T}]^T$ is converging. With notations defined above, we can rewrite $\mathbf{u}^t = -(\mathcal{H}^t)^{-1} \mathbf{g}^t(\mathbf{0})$ as

$$\mathbf{u}^t = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{w}^{t-1}. \quad (5.51)$$

The vector $\mathbf{w}^t$ is updated via

$$\mathbf{w}^t = \tilde{\mathbf{g}}(\mathbf{P} \mathbf{w}^{t-1} + \mathbf{C} \mathbf{s}^{t-1}) = \tilde{\mathbf{g}}([\mathbf{P} \quad \mathbf{C}] \boldsymbol{\theta}^{t-1}), \quad (5.52)$$

where $\tilde{\mathbf{g}}(\mathbf{v}) = \mathbf{D}^{\frac{1}{2}} \mathbf{g}(\mathbf{D}^{-\frac{1}{2}} \mathbf{v})$. The update of $\mathbf{s}^t$ can be written as

$$\begin{aligned} \mathbf{s}^t &= \mathbf{s}^{t-1} + \mathbf{H}(\mathbf{w}^t - \mathbf{B} \mathbf{u}^t) \\ &= \mathbf{s}^{t-1} + \mathbf{H}[\tilde{\mathbf{g}}([\mathbf{P} \quad \mathbf{C}] \boldsymbol{\theta}^{t-1}) - \mathbf{P} \mathbf{w}^{t-1}]. \end{aligned} \quad (5.53)$$

Combining (5.52) and (5.53), we obtain the update equation for $\boldsymbol{\theta}^t$,

$$\boldsymbol{\theta}^t = \mathbf{h}(\boldsymbol{\theta}^{t-1}) = \begin{bmatrix} \mathbf{I} \\ \mathbf{H} \end{bmatrix} \tilde{\mathbf{g}}([\mathbf{P} \ \mathbf{C}] \boldsymbol{\theta}^{t-1}) + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{H}\mathbf{P} & \mathbf{I} \end{bmatrix} \boldsymbol{\theta}^{t-1}, \quad (5.54)$$

where $\mathbf{h}(\boldsymbol{\theta}^{t-1})$ denotes the update function. We get for the Jacobian $\tilde{\mathbf{g}}^t = \tilde{\mathbf{g}}'([\mathbf{P} \ \mathbf{C}] \boldsymbol{\theta}^{t-1}) = \mathbf{D}^{\frac{1}{2}} \mathbf{g}^t \mathbf{D}^{-\frac{1}{2}} = \mathbf{g}^t$ which is diagonal since $\mathbf{g}(\cdot)$ is an elementwise function. As mentioned in [42], $\mathbf{g}^t$ is a positive semi-definite diagonal matrix with all elements smaller than 1. Furthermore, in the Gaussian case $\mathbf{g}^t$ is a constant matrix. The Jacobian of $\mathbf{h}(\boldsymbol{\theta}^{t-1})$ is given by

$$\begin{aligned} \mathbf{h}^t &= \begin{bmatrix} \mathbf{I} \\ \mathbf{H} \end{bmatrix} \mathbf{g}^t [\mathbf{P} \ \mathbf{C}] + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ -\mathbf{H}\mathbf{P} & \mathbf{I} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{g}^t \mathbf{P} & \mathbf{g}^t \mathbf{C} \\ \mathbf{H}(\mathbf{g}^t - \mathbf{I})\mathbf{P} & \mathbf{H}\mathbf{g}^t \mathbf{C} + \mathbf{I} \end{bmatrix}. \end{aligned} \quad (5.55)$$

We calculate the terms $\mathbf{H}\mathbf{g}^t \mathbf{C}$ and $\mathbf{H}(\mathbf{g}^t - \mathbf{I})$, which leads to

$$\mathbf{H}\mathbf{g}^t \mathbf{C} = -\mathbf{g}_z^t, \quad \mathbf{H}(\mathbf{g}^t - \mathbf{I}) = (\mathbf{g}_z^t - \mathbf{I})\mathbf{H}. \quad (5.56)$$

Hence, the Jacobian $\mathbf{h}^t$ becomes

$$\mathbf{h}^t = \begin{bmatrix} \mathbf{g}^t \mathbf{P} & \mathbf{g}^t \mathbf{C} \\ (\mathbf{g}_p^t - \mathbf{I})\mathbf{H}\mathbf{P} & \mathbf{I} - \mathbf{g}_z^t \end{bmatrix} = \begin{bmatrix} \mathbf{g}^t & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \mathbf{g}_z^t \end{bmatrix} \begin{bmatrix} \mathbf{P} & \mathbf{C} \\ -\mathbf{H}\mathbf{P} & \mathbf{I} \end{bmatrix}. \quad (5.57)$$

Since all the elements of $\mathbf{g}^t$ range from 0 to 1,

$$\mathbf{0} \leq \begin{bmatrix} \mathbf{g}^t & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \mathbf{g}_z^t \end{bmatrix} \leq \max\{\|\mathbf{g}^t\|_\infty, 1 - \|\mathbf{g}_z^t\|_\infty\} \mathbf{I} < \mathbf{I}. \quad (5.58)$$

We calculate the eigenvalues of the second matrix at the right of (5.57) via

$$\det \begin{bmatrix} \lambda \mathbf{I} - \mathbf{P} & -\mathbf{C} \\ \mathbf{H}\mathbf{P} & \lambda \mathbf{I} - \mathbf{I} \end{bmatrix} = 0. \quad (5.59)$$

For the determinant of block matrices, we have

$$\det \begin{bmatrix} \lambda \mathbf{I} - \mathbf{P} & -\mathbf{C} \\ \mathbf{H}\mathbf{P} & \lambda \mathbf{I} - \mathbf{I} \end{bmatrix} = \det(\lambda \mathbf{I} - \mathbf{P}) \det[\lambda \mathbf{I} - \mathbf{I} + \mathbf{H}\mathbf{P}(\lambda \mathbf{I} - \mathbf{P})^{-1} \mathbf{C}]. \quad (5.60)$$

By the matrix inverse lemma and the definition of $\mathbf{P}$, we get

$$(\lambda \mathbf{I} - \mathbf{P})^{-1} = \frac{\mathbf{I}}{\lambda} - \frac{\mathbf{P}}{\lambda - \lambda^2}. \quad (5.61)$$

Note that from the definition in (5.49), $\mathbf{P}\mathbf{C} = \mathbf{B}(\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{C} = \mathbf{0}$. Hence (5.60) becomes

$$\det \begin{bmatrix} \lambda \mathbf{I} - \mathbf{P} & -\mathbf{C} \\ \mathbf{H}\mathbf{P} & \lambda \mathbf{I} - \mathbf{I} \end{bmatrix} = \det(\lambda \mathbf{I} - \mathbf{P}) \det(\lambda \mathbf{I} - \mathbf{I}). \quad (5.62)$$

Set this determinant to 0 and solve for λ , we find that the eigenvalues are 0 and 1. Therefore, the update operation $\mathbf{h}(\boldsymbol{\theta})$ is contractive.

5.8 Simulation Results

For the simulations, we set the SNR to 20dB, and the system dimensions to $M \times N = 512 \times 1024$. We consider a Gaussian setting with white noise and $\mathbf{x}$ is drawn from an n.i.i.d. Gaussian distribution with zero mean and variance profile $\sigma_{x_i}^2 = 0.991^{i-1}$, $i = 1, \dots, N$. For $\mathbf{A}$, we follow the setup in [33]. Namely first $\mathbf{A}$ gets generated as i.i.d. zero mean Gaussian, its SVD gets computed and the singular values $\{s_1 \geq \dots \geq s_M\}$ are changed to a geometric series with a specific condition number $\frac{s_1}{s_M}$. We compare the results to the LMMSE estimator. Fig. 5.1 illustrates the difference between the $\hat{\mathbf{x}}^t$ and the LMMSE $\hat{\mathbf{x}}_{\text{MMSE}}$, whereas Fig. 5.2 compares the difference between $\boldsymbol{\tau}_x^t$ and $\boldsymbol{\tau}_{\text{MMSE}}$. The "normalization" mentioned in the captions refers to division by N . These simulations show that the AMBGAMP algorithm continues to work in unrealistically severe scenarios (in which AMP diverges).

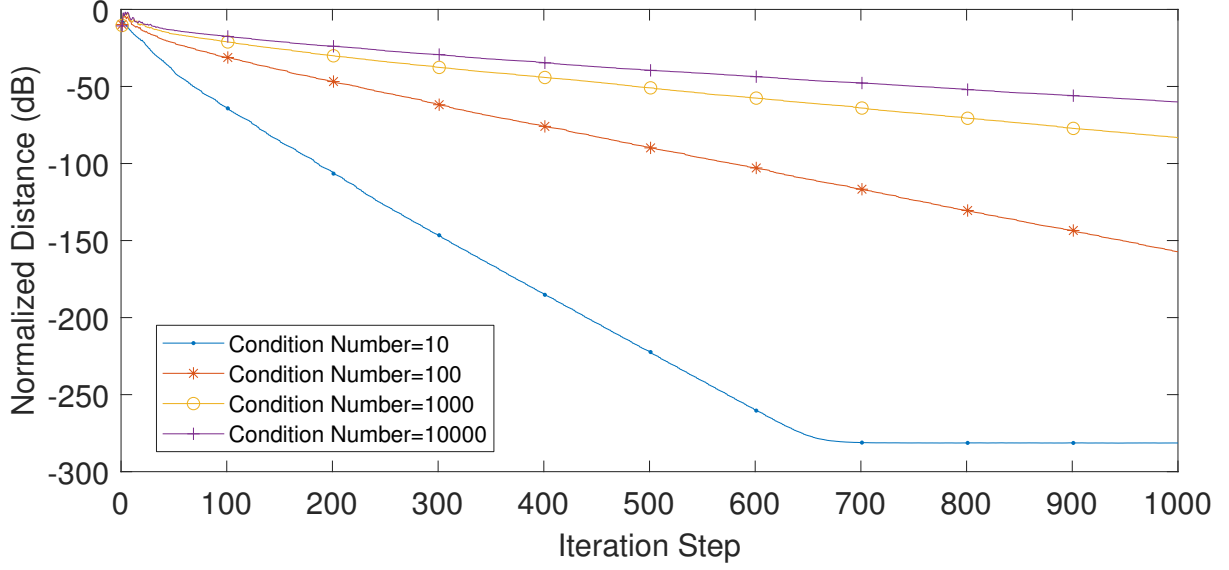


Figure 5.1: Normalized l_2 norm of $\hat{\mathbf{x}}^t - \hat{\mathbf{x}}_{\text{MMSE}}$

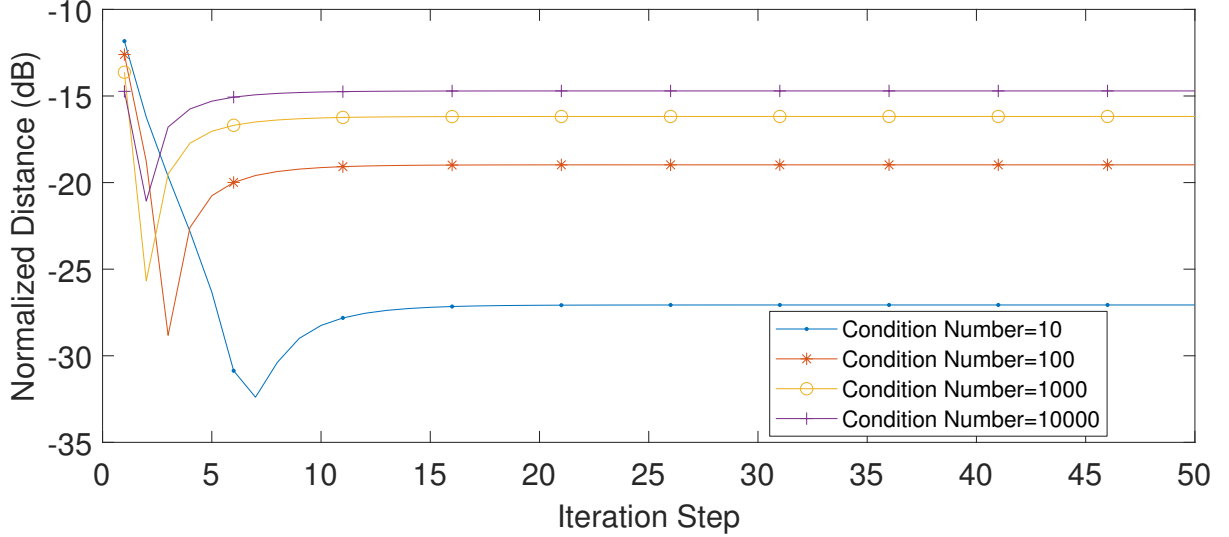


Figure 5.2: Normalized l_1 norm of $\boldsymbol{\tau}_x^t - \boldsymbol{\tau}_{\text{MMSE}}$

5.9 Concluding Remarks

We propose a convergent version of GAMP, AMBGAMP, which applies alternating minimization to an augmented Lagrangian of a large system limit of the Bethe free Energy (BFE). AMBGAMP can be interpreted as applying a simplified ADMM to the BFE, with a constrained Lagrange multiplier parametrization for the mean constraint, and a quadratic optimization subproblem being solved by a gradient update with line search. The ADMM is complemented with a fixed point iteration for the variance constraint.

Part II

Bilinear Estimation

Chapter 6

Bilinear Hybrid Expectation Maximization and Expectation Propagation for Semi-Blind Channel Estimation

6.1 Introduction

One of the unique features of Cell-Free (CF) Massive MIMO (MaMIMO) networks is the absence of traditional cellular boundaries, which introduces a critical challenge: the potential for a high number of user terminals (UTs) in a given area to exceed the length of the pilot sequence. This imbalance leads to what is known as pilot contamination. Addressing this issue, Semi-Blind approaches have been explored as a viable solution to mitigate the effects of pilot contamination, as detailed in the work of Gholami et al. [50].

In the context of Bayesian inference, the Semi-Blind approach is modeled as a bilinear inference problem, where the Access Points (APs) must concurrently estimate both the channel state information (CSI) and the user signals. [51] have demonstrated that transitioning from the deterministic domain to the Bayesian domain results in significant gains. The challenge in addressing the bilinear inference problem primarily arises from obtaining the closed-form solution of the marginal likelihood of the observed data given the channel coefficient. To circumvent this issue, researchers have been exploring iterative approximation techniques such as expectation maximization (EM) and expectation propagation (EP).

6.1.1 Prior Works

Expectation-maximization

Expectation-maximization can be viewed as a special case of the Majorize-Minimization (MM) algorithm. In [52], the authors contrasted two variants of the EM approach within the context of semi-blind channel estimation: one incorporates prior information, while the other disregards it. This is essentially a comparison between Bayesian and deterministic approaches. The authors of [53] proposed a space-alternating generalized expectation maximization (SAGE) algorithm to handle the lack of prior information. To tackle the challenges posed by systems with a more generalized prior distribution for input signals,

[54] explored the combination of EM-based on a Gaussian mixture prior.

Expectation-propagation

Message-passing algorithms play a critical role in the inference problems. Notably, one of the most potent message-passing algorithms is Expectation Propagation (EP), proposed by Minka [9]. Based on EP rules and further approximations, bilinear G-AMP (BiG-AMP) was proposed in [23]. This algorithm is also capable of automatically estimating the prior information by employing the EM technique to estimate the hyperparameters. In the scenario of Cell-Free MaMIMO, the base station number is typically very high. Hence, to distribute the computational load between the CPU and all Base Stations, a distributed version of Expectation Propagation, termed "bilinear EP for joint channel estimation and data detection (JCD)," was proposed in [2]. The close relation between EP and Bethe Free Energy (BFE) was studied in [6], which shows that they share the same fixed points. Given that the primary concern with EP is the uncertainty of convergence, a more robust solution can be achieved by directly analyzing the optimization of the BFE. A loop-free EP was proposed in [21] based on approximate BFE optimization.

In [20], the authors proposed variable level expectation propagation (VL-EP) for semi-blind channel estimation. They further explored the relation between EM-based methods and VL-EP and showed the close relation between VL-EP and EM-based algorithms in semi-blind channel estimation.

6.1.2 Main Contribution

In our study, we employ the expression for EP as outlined by [10] to develop an approximate probabilistic model for the bilinear system. Leveraging this model, we have identified that EM-based methods, as described in [20], can be improved by incorporating moment matching for the approximated posteriors of the per-user received signals. Consequently, we propose a hybrid EM-EP algorithm designed to correctly manage the posterior distribution of the per-user received signals. To demonstrate the effectiveness of our method, we present simulation results comparing our proposed algorithm with the VL-EP algorithm and the Gaussian input semi-blind Cramer-Rao Bound (CRB), thus providing evidence of the superiority of our approach.

6.2 System Model

We examine an uplink semi-blind signal model described as follows:

$$\mathbf{Y} = [\mathbf{Y}_p \quad \mathbf{Y}_d] = \mathbf{H} [\mathbf{X}_p^T \quad \mathbf{X}^T] + [\mathbf{V}_p \quad \mathbf{V}] \in \mathcal{C}^{M \times (P+T)}, \quad (6.1)$$

where $\mathbf{H}$ represents the channel matrix, unknown and modeled as an independent and identically distributed (i.i.d.) random matrix of size $M \times K$. Each column follows the distribution $\mathbf{h}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{\Xi}_{\mathbf{h}_k})$. The input signal is composed of a pilot part $\mathbf{X}_p \in \mathcal{C}^{P \times K}$ and a data part $\mathbf{X} \in \mathcal{C}^{T \times K}$. It's assumed that all data symbols in $\mathbf{X}$ are independently drawn from a Gaussian distribution with power σ_x^2 .

We define $\mathbf{V}_p$ and $\mathbf{V}$ as the Additive White Gaussian Noise (AWGN) at the APs. Each element within these noise matrices is assumed to be independently drawn from a Gaussian distribution with power σ_v^2 .

6.2.1 Orthogonal Pilot

When orthogonal pilot is used, we can correlate $\mathbf{Y}_p$ with each pilot sequence $\mathbf{x}_{p,g}^*$ to obtain an equivalent observation of the pilot $\mathbf{y}_g$ such that

$$\mathbf{y}_{p,g} = \mathbf{Y}_p \mathbf{x}_{p,g}^* = P\sigma_x^2 \sum_{k \in G_g} \mathbf{h}_k + \mathbf{v}_{p,g}, \quad (6.2)$$

where $\mathbf{v}_{p,g} = \mathbf{V}_p \mathbf{x}_{p,g}^* \sim \mathcal{N}(\mathbf{v}_{p,g} | \mathbf{0}, \sigma_x^2 \sigma_v^2 P \mathbf{I})$ and G_g denotes the group of users using the g -th pilot sequence. Since one UT k can only send one pilot sequence, the user groups $G_g \cap G_{g'} = \emptyset$ if $g \neq g'$. Due to orthogonal pilots, we also have $\mathbb{E}[\mathbf{y}_{p,g} \mathbf{y}_{p,g'}] = \mathbf{0}$ if $g \neq g'$.

6.3 EP Approximation

We treat the vectors $\mathbf{h}_k$ and $\mathbf{x}_k$ as atomic variables and utilize an EP-like algorithm to infer their posterior means and covariances.

The joint probability of the underlying system model is

$$\begin{aligned} p(\mathbf{Y}_p, \mathbf{Y}_d, \mathbf{h}_{\{k\}}, \mathbf{x}_{\{k\}}) &= p(\mathbf{y}_{p,\{g\}}, \mathbf{Y}_d, \mathbf{h}_{\{k\}}, \mathbf{x}_{\{k\}}) \\ &= p(\mathbf{Y}_d | \mathbf{h}_{\{k\}}, \mathbf{x}_{\{k\}}) \prod_g p(\mathbf{y}_{p,g}, \mathbf{H}_g) \prod_k p(\mathbf{x}_k), \end{aligned} \quad (6.3)$$

where we use $\{\cdot\}$ to iterate all the indices, and the first equal sign is obtained because of orthogonal pilot.

According to EP, we approximate the joint pdf (6.3) by $b(\mathbf{H}, \mathbf{X})$

$$\begin{aligned} p(\mathbf{y}_{p,\{g\}}, \mathbf{Y}_d, \mathbf{h}_{\{k\}}, \mathbf{x}_{\{k\}}) &\simeq b(\mathbf{H}, \mathbf{X}) \\ &: \propto q_{\mathbf{H}, \mathbf{X}}(\mathbf{h}_{\{k\}}, \mathbf{x}_{\{k\}}) \prod_g q_{\mathbf{H}_g}(\mathbf{h}_{\{k \in G_g\}}) \prod_k p(\mathbf{x}_k) \\ &= \mu_{q_{\mathbf{H}, \mathbf{X}}; \mathbf{h}_k}(\mathbf{h}_k) \mu_{q_{\mathbf{H}, \mathbf{X}}; \mathbf{x}_k}(\mathbf{x}_k) \mu_{q_{\mathbf{H}}; \mathbf{h}_k}(\mathbf{h}_k) p(\mathbf{x}_k), \end{aligned} \quad (6.4)$$

where we further assume that $q_{\mathbf{H}, \mathbf{X}}$ and $q_{\mathbf{H}_g}$ can be factorized at variable level,

$$\begin{aligned} q_{\mathbf{H}, \mathbf{X}}(\mathbf{H}, \mathbf{X}) &= \prod_k \mu_{q_{\mathbf{H}, \mathbf{X}}; \mathbf{h}_k}(\mathbf{h}_k) \mu_{q_{\mathbf{H}, \mathbf{X}}; \mathbf{x}_k}(\mathbf{x}_k) \\ q_{\mathbf{H}_g}(\mathbf{H}_g) &= \prod_{k \in G_g} \mu_{q_{\mathbf{H}}; \mathbf{h}_k}(\mathbf{h}_k). \end{aligned} \quad (6.5)$$

From this point, we can see that refining each approximated factor $q_{\mathbf{H}, \mathbf{X}}$ and $q_{\mathbf{H}_g}$ is equivalent to refining the variable level factors $\mu_{q_{\mathbf{H}, \mathbf{X}}; \mathbf{h}_k}$, $\mu_{q_{\mathbf{H}, \mathbf{X}}; \mathbf{x}_k}$ and $\mu_{q_{\mathbf{H}}; \mathbf{h}_k}$ of the target factors in parallel.

6.4 EM-EP

In EP algorithm, we refine the approximated factors $q_{\mathbf{H}, \mathbf{X}}$ and $q_{\mathbf{H}_g}$ iteratively. We first look at the refinement for the bilinear factor $q_{\mathbf{H}, \mathbf{X}}$.

The KLD objective function for refining $q_{\mathbf{H}, \mathbf{X}}$ is

$$\text{KLD}[b_{q_{\mathbf{H}, \mathbf{X}}}(\mathbf{H}, \mathbf{X}) \| b(\mathbf{H}, \mathbf{X})], \quad (6.6)$$

where

$$b_{q_{\mathbf{H},\mathbf{X}}}(\mathbf{H}, \mathbf{X}) \propto p(\mathbf{Y}_d|\mathbf{H}, \mathbf{X}) \prod_g q_{\mathbf{H}_g}(\mathbf{h}_{\{k \in G_g\}}) \prod_k p(\mathbf{x}_{d,k}) \quad (6.7)$$

is the belief at factor $p(\mathbf{Y}_d|\mathbf{H}, \mathbf{X})$. Substitute (6.4) and ignore the terms that are irrelevant to $q_{\mathbf{H},\mathbf{X}}$, the KLD objective function becomes

$$\begin{aligned} \text{KLD}[b_{q_{\mathbf{H},\mathbf{X}}}(\mathbf{H}, \mathbf{X}) \| b(\mathbf{H}, \mathbf{X})] \\ = - \sum_k \int b_{q_{\mathbf{H},\mathbf{X}}}(\mathbf{H}, \mathbf{X}) \ln \mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{h}_k}(\mathbf{h}_k) \mu_{q_{\mathbf{H}};\mathbf{h}_k}(\mathbf{h}_k) d\mathbf{H} d\mathbf{X} \\ - \sum_k \int b_{q_{\mathbf{H},\mathbf{X}}}(\mathbf{H}, \mathbf{X}) \ln \mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{x}_k}(\mathbf{x}_k) p(\mathbf{x}_k) d\mathbf{H} d\mathbf{X} + \ln Z_{b_{\mathbf{H},\mathbf{X}}}, \end{aligned} \quad (6.8)$$

where $Z_{b_{\mathbf{H},\mathbf{X}}}$ is the normalization factor corresponding to (6.4), i.e., $Z_{b_{\mathbf{H},\mathbf{X}}} = Z_{b_{\mathbf{h}_k}} Z_{b_{\mathbf{x}_k}}$ with

$$\begin{aligned} Z_{b_{\mathbf{h}_k}} &= \int \mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{h}_k}(\mathbf{h}_k) \mu_{q_{\mathbf{H}};\mathbf{h}_k}(\mathbf{h}_k) d\mathbf{h}_k \\ Z_{b_{\mathbf{x}_k}} &= \int \mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{x}_k}(\mathbf{x}_k) p(\mathbf{x}_k) d\mathbf{x}_k. \end{aligned} \quad (6.9)$$

At this point, we observe that all the variable level factors $\mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{h}_k}(\mathbf{h}_k)$ and $\mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{x}_k}(\mathbf{x}_k)$ of $q_{\mathbf{H},\mathbf{X}}$ can be decoupled. Therefore, refining the factor $q_{\mathbf{H},\mathbf{X}}$ based on (6.8) is equivalent to refining all the variable level factors $\mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{h}_k}(\mathbf{h}_k)$ and $\mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{x}_k}(\mathbf{x}_k)$ in parallel. We look at the refinement of $\mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{h}_k}(\mathbf{h}_k)$. We omit the terms that does not contain $\mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{h}_k}(\mathbf{h}_k)$ from (6.8), and the objective function for updating it is

$$\begin{aligned} \text{KLD}[b_{q_{\mathbf{H},\mathbf{X}}}(\mathbf{H}, \mathbf{X}) \| b(\mathbf{H}, \mathbf{X})] \\ = \text{KLD} \left[b_{q_{\mathbf{H},\mathbf{X}}}(\mathbf{h}_k) \left\| \frac{\mu_{q_{\mathbf{H},\mathbf{X}};\mathbf{h}_k}(\mathbf{h}_k) \mu_{q_{\mathbf{H}};\mathbf{h}_k}(\mathbf{h}_k)}{Z_{b_{\mathbf{h}_k}}} \right\| \right] + c, \end{aligned} \quad (6.10)$$

where marginalized belief is denoted as $b_{q_{\mathbf{H},\mathbf{X}}}(\mathbf{h}_k) = \int b_{q_{\mathbf{H},\mathbf{X}}}(\mathbf{H}, \mathbf{X}) d\mathbf{h}_{\bar{k}} d\mathbf{X}$, with the notation $\mathbf{h}_{\bar{k}}$ denoting the set of channel coefficient vectors $\forall i \neq k, \mathbf{h}_i$. The marginal belief $b_{q_{\mathbf{H},\mathbf{X}}}$ is computed as

$$\begin{aligned} b_{q_{\mathbf{H},\mathbf{X}}}(\mathbf{h}_k) &\propto \int \mu_{q_{\mathbf{H}};\mathbf{h}_i}(\mathbf{h}_i) p(\mathbf{x}_i) p(\mathbf{Y}_d|\mathbf{h}_k \mathbf{x}_k^T + \sum_{i \neq k} \mathbf{h}_i \mathbf{x}_i^T) \\ &\cdot \prod_{i \neq k} \mu_{q_{\mathbf{H}};\mathbf{h}_i}(\mathbf{h}_i) p(\mathbf{x}_i) d\mathbf{h}_{\bar{k}} d\mathbf{X}. \end{aligned} \quad (6.11)$$

Due to Central Limit Theory (CLT), we can approximate the interference term $\sum_{i \neq k} \mathbf{h}_i \mathbf{x}_i^T$ to a Gaussian, where the component variables $\mathbf{h}_i$ and $\mathbf{x}_i$ follows $\mathbf{h}_i \sim \mu_{q_{\mathbf{H}};\mathbf{h}_i}(\mathbf{h}_i)$ and $\mathbf{x}_i \sim p(\mathbf{x}_i)$. We define a Gaussian random matrix $\mathbf{Z}_k \simeq \sum_{i \neq k} \mathbf{h}_i \mathbf{x}_i^T$ and define vector $\mathbf{z}_k = \text{vec}(\mathbf{Z}_k)$ to approximate the interference,

$$\mathbf{z}_k \sim \mathcal{N}(\mathbf{z}_k | \mathbf{m}_{\mathbf{z}_k}, \mathbf{C}_{\mathbf{z}_k}), \quad (6.12)$$

where $\mathbf{m}_{\mathbf{z}_k}$ and $\mathbf{C}_{\mathbf{z}_k}$ equal to the mean and covariance matrix of the vectorized interference $\text{vec}(\sum_{i \neq k} \mathbf{h}_i \mathbf{x}_i^T)$,

$$\begin{aligned} \mathbf{m}_{\mathbf{z}_k} &= \mathbf{0} \\ \mathbf{C}_{\mathbf{z}_k} &= \sigma_x^2 \left[\mathbf{I}_T \otimes \sum_{i \neq k} (\mathbf{C}_{q_{\mathbf{H}};\mathbf{h}_i} + \mathbf{m}_{q_{\mathbf{H}};\mathbf{h}_i} \mathbf{m}_{q_{\mathbf{H}};\mathbf{h}_i}^H) \right] \end{aligned} \quad (6.13)$$

With this CLT approximation, (6.11) becomes

$$\begin{aligned} b_{q_{H,X}}(\mathbf{h}_k) &\simeq \int \mu_{q_{H;\mathbf{h}_k}}(\mathbf{h}_k) p(\mathbf{x}_k) p(\mathbf{Y}_d | \mathbf{h}_k \mathbf{x}_k^T, \mathbf{C}_{\mathbf{z}_k}) d\mathbf{x}_k \\ &= \mu_{q_{H;\mathbf{h}_k}}(\mathbf{h}_k) p(\mathbf{Y}_d | \mathbf{h}_k, \mathbf{C}_{\mathbf{z}_k}) \end{aligned} \quad (6.14)$$

where $p(\mathbf{Y}_d | \mathbf{h}_k \mathbf{x}_k^T, \mathbf{C}_{\mathbf{z}_k})$ is defined as

$$\begin{aligned} p(\mathbf{Y}_d | \mathbf{h}_k \mathbf{x}_k^T, \mathbf{C}_{\mathbf{z}_k}) &= \int p(\mathbf{Y}_d | \mathbf{h}_k \mathbf{x}_k^T + \mathbf{Z}_k) \mathcal{N}(\mathbf{z}_k | \mathbf{m}_{\mathbf{z}_k}, \mathbf{C}_{\mathbf{z}_k}) d\mathbf{z}_k \\ &= \mathcal{N}(\mathbf{y}_d | \text{vec}(\mathbf{h}_k \mathbf{x}_k^T), \mathbf{I}_T \otimes \mathbf{C}_{\tilde{\mathbf{v}}_k}), \end{aligned}$$

with the notation $\mathbf{y}_d = \text{vec}(\mathbf{Y}_d)$ and noise plus interference covariance matrix

$$\mathbf{C}_{\tilde{\mathbf{v}}_k} = \sigma_v^2 \mathbf{I} + \sigma_x^2 \sum_{i \neq k} (\mathbf{C}_{q_{H;\mathbf{h}_i}} + \mathbf{m}_{q_{H;\mathbf{h}_i}} \mathbf{m}_{q_{H;\mathbf{h}_i}}^H)$$

To obtain the posterior mean and covariance of $b_{q_{H,X}}$, we consider an Expectation Maximization approach and approximate by a surrogate majorization function. The negated log-likelihood of $p(\mathbf{Y}_d | \mathbf{h}_k, \mathbf{C}_{\tilde{\mathbf{v}}_k})$ is

$$l_{\mathbf{h}_k}(\mathbf{h}_k) = -\ln p(\mathbf{Y}_d | \mathbf{h}_k, \mathbf{C}_{\tilde{\mathbf{v}}_k}). \quad (6.15)$$

Since we have $p(\mathbf{x}_k) = \mathcal{N}(\mathbf{x}_k | \mathbf{0}, \sigma_x^2 \mathbf{I})$ and the sum $\mathbf{v}_d + \mathbf{z}_k$ has a block diagonal covariance matrix, the marginalized log-likelihood function is circular symmetric, i.e. $\forall \varphi, l_{\mathbf{h}_k}(\mathbf{h}_k) = l_{\mathbf{h}_k}(\mathbf{h}_k e^{j\varphi})$. We construct a surrogate function,

$$\begin{aligned} u_{\mathbf{h}_k}(\mathbf{h}_k | \hat{\mathbf{h}}_k) \\ = l_{\mathbf{h}_k}(\hat{\mathbf{h}}_k) - \mathbb{E}_{\mathbf{x}_k | \mathbf{Y}_d, \hat{\mathbf{h}}_k, \mathbf{C}_{\tilde{\mathbf{v}}_k}} \left[\ln \left(\frac{p(\mathbf{Y}_d, \mathbf{x}_k | \mathbf{h}_k, \mathbf{C}_{\tilde{\mathbf{v}}_k})}{p(\mathbf{Y}_d, \mathbf{x}_k | \hat{\mathbf{h}}_k, \mathbf{C}_{\tilde{\mathbf{v}}_k})} \right) \right], \end{aligned}$$

where $\hat{\mathbf{h}}_k$ is the posterior estimate from previous iteration. One can verify by Jensen's inequality that $u_{\mathbf{h}_k}(\mathbf{h}_k | \hat{\mathbf{h}}_k) \geq l_{\mathbf{h}_k}(\mathbf{h}_k)$, where equality is reached at $\mathbf{h}_k = \hat{\mathbf{h}}_k$. This implies that following relation of the log belief pdf $-\ln(b_{q_{H,X}}) \leq u_{\mathbf{h}_k}(\mathbf{h}_k | \hat{\mathbf{h}}_k) - \ln(\mu_{q_{H;\mathbf{h}_k}}(\mathbf{h}_k))$. We approximate the belief $b_{q_{H,X}}(\mathbf{h}_k)$ by the minorization function $\exp[-u_{\mathbf{h}_k}(\mathbf{h}_k | \hat{\mathbf{h}}_k) + \ln(\mu_{q_{H;\mathbf{h}_k}}(\mathbf{h}_k))]$ which we will show to be a Gaussian. Since $\mu_{q_{H;\mathbf{h}_k}}(\mathbf{h}_k)$ is already Gaussian, we focus on the term $u_{\mathbf{h}_k}(\mathbf{h}_k | \hat{\mathbf{h}}_k)$. By omitting constant terms that do not involve $\mathbf{h}_k$, we derive an effective surrogate function of $l_{\mathbf{h}_k}(\mathbf{h}_k)$

$$\begin{aligned} \bar{u}_{\mathbf{h}_k}(\mathbf{h}_k | \hat{\mathbf{h}}_k) &= -\mathbb{E}_{\mathbf{x}_k | \mathbf{Y}_d, \hat{\mathbf{h}}_k, \mathbf{C}_{\tilde{\mathbf{v}}_k}} [\ln(p(\mathbf{Y}_d | \mathbf{x}_k, \mathbf{h}_k, \mathbf{C}_{\tilde{\mathbf{v}}_k}))] \\ &= \|\mathbf{y}_d - (\hat{\mathbf{x}}_k \otimes \mathbf{I}_M) \mathbf{h}_k\|_{(\mathbf{I}_T \otimes \mathbf{C}_{\tilde{\mathbf{v}}_k})^{-1}}^2 + \text{tr}[\mathbf{C}_{\hat{\mathbf{x}}}] \|\mathbf{h}_k\|_{\mathbf{C}_{\tilde{\mathbf{v}}_k}^{-1}}^2, \end{aligned}$$

where $\hat{\mathbf{x}}_k$ and $\mathbf{C}_{\hat{\mathbf{x}}}$ are the mean and covariance matrix of the posterior $p(\mathbf{x}_k | \mathbf{Y}_k, \hat{\mathbf{h}}_k, \mathbf{C}_{\mathbf{z}_k})$. These statistics can be obtained via LMMSE as

$$\begin{aligned} \tau_{\mathbf{x}_k} &= \left(\hat{\mathbf{h}}_k^H \mathbf{C}_{\tilde{\mathbf{v}}_k}^{-1} \hat{\mathbf{h}}_k + \sigma_x^{-2} \right)^{-1} \\ \mathbf{C}_{\hat{\mathbf{x}}_k} &= \tau_{\mathbf{x}_k} \mathbf{I}_T \\ \hat{\mathbf{x}}_k^T &= \tau_{\mathbf{x}_k} \hat{\mathbf{h}}_k^H \mathbf{C}_{\tilde{\mathbf{v}}_k}^{-1} \mathbf{Y}_d. \end{aligned} \quad (6.16)$$

By combining $\bar{u}(\mathbf{h}_k|\hat{\mathbf{h}}_k)$ with the non-constant terms in $-\ln(\mu_{q_{\mathbf{H}};\mathbf{h}_k}(\mathbf{h}_k))$, we obtain an effective surrogate function for approximating $-\ln(b_{q_{\mathbf{H}},\mathbf{X}}(\mathbf{h}_k))$,

$$\begin{aligned} -\ln(b_{q_{\mathbf{H}},\mathbf{X}}(\mathbf{h}_k)) &\leq \bar{u}(\mathbf{h}_k|\hat{\mathbf{h}}_k) - \ln(\mu_{q_{\mathbf{H}};\mathbf{h}_k}(\mathbf{h}_k)) + c \\ &= \|\mathbf{y}_d - (\hat{\mathbf{x}}_k \otimes \mathbf{I}_M)\mathbf{h}_k\|_{(\mathbf{I}_T \otimes \mathbf{C}_{\hat{\mathbf{v}}_k})^{-1} + \text{tr}[\mathbf{C}_{\hat{\mathbf{x}}}] \|\mathbf{h}_k\|_{\mathbf{C}_{\hat{\mathbf{v}}_k}^{-1}}^2 \\ &\quad + \|\mathbf{h}_k - \mathbf{m}_{q_{\mathbf{H}};\mathbf{h}_k}\|_{\mathbf{C}_{q_{\mathbf{H}};\mathbf{h}_k}^{-1}}^2. \end{aligned} \quad (6.17)$$

It can be identified as an Gaussian with mean and covariance

$$\begin{aligned} \mathbf{C}_{\mathbf{h}_k} &= \left[(\hat{\mathbf{x}}_k^H \hat{\mathbf{x}}_k + T\tau_{\hat{\mathbf{x}}_k}) \mathbf{C}_{\hat{\mathbf{v}}_k}^{-1} + \mathbf{C}_{q_{\mathbf{H}};\mathbf{h}_k}^{-1} \right]^{-1} \\ \mathbf{m}_{\mathbf{h}_k} &= \mathbf{C}_{\mathbf{h}_k} \left(\mathbf{C}_{\hat{\mathbf{v}}_k}^{-1} \mathbf{Y}_d \hat{\mathbf{x}}_k^* + \mathbf{C}_{q_{\mathbf{H}};\mathbf{h}_k}^{-1} \mathbf{m}_{q_{\mathbf{H}};\mathbf{h}_k} \right) \end{aligned} \quad (6.18)$$

Optimizing the KLD (6.10) and substituting $b_{q_{\mathbf{H}},\mathbf{X}}(\mathbf{h}_k) = \mathcal{N}(\mathbf{h}_k|\mathbf{m}_{\mathbf{h}_k}, \mathbf{C}_{\mathbf{h}_k})$, we obtain the update $\mu_{q_{\mathbf{H}},\mathbf{X};\mathbf{h}_k}(\mathbf{h}_k) = \frac{\mathcal{N}(\mathbf{h}_k|\mathbf{m}_{\mathbf{h}_k}, \mathbf{C}_{\mathbf{h}_k})}{\mu_{q_{\mathbf{H}};\mathbf{h}_k}(\mathbf{h}_k)}$, where the quotient is a Gaussian with mean and covariance matrix given by

$$\begin{aligned} \mathbf{C}_{q_{\mathbf{H}},\mathbf{X};\mathbf{h}_k} &= \frac{\mathbf{C}_{\hat{\mathbf{v}}_k}}{\hat{\mathbf{x}}_k^H \hat{\mathbf{x}}_k + T\tau_{\hat{\mathbf{x}}_k}} \\ \mathbf{m}_{q_{\mathbf{H}},\mathbf{X};\mathbf{h}_k} &= \frac{\mathbf{Y}_d \hat{\mathbf{x}}_k^*}{\hat{\mathbf{x}}_k^H \hat{\mathbf{x}}_k + T\tau_{\hat{\mathbf{x}}_k}}. \end{aligned} \quad (6.19)$$

To refine the approximate $q_{\mathbf{H}_g}(\mathbf{H}_g) = \prod_{k \in G_g} \mu_{q_{\mathbf{H}};\mathbf{h}_k}(\mathbf{h}_k)$, we optimize the KLD objective function

$$\text{KLD}[b_{q_{\mathbf{H}_g}}(\mathbf{H}, \mathbf{X}) \| b(\mathbf{H}, \mathbf{X})], \quad (6.20)$$

where the belief

$$b_{q_{\mathbf{H}_g}}(\mathbf{H}, \mathbf{X}) = p(\mathbf{y}_{p,g}, \mathbf{H}_g) q_{\mathbf{H},\mathbf{X}}(\mathbf{H}, \mathbf{X}) \prod_{j \neq g} q_{\mathbf{H}_j}(\mathbf{H}_j) \prod_k p(\mathbf{x}_k).$$

Substituting $q_{\mathbf{H},\mathbf{X}}$, $q_{\mathbf{H}_j}$ with (6.5), we observe that optimizing $q_{\mathbf{H}_g}$ in (6.20) is equivalent to updating the all the variable level factors $\forall k \in G_g, \mu_{q_{\mathbf{H}};\mathbf{h}_k}$ in parallel, i.e.

$$\begin{aligned} &\text{KLD}[b_{q_{\mathbf{H}_g}}(\mathbf{H}, \mathbf{X}) \| b(\mathbf{H}, \mathbf{X})] \\ &= \sum_{k \in G_g} \text{KLD} \left[b_{q_{\mathbf{H}_g}}(\mathbf{H}_g) \left\| \frac{\mu_{q_{\mathbf{H}},\mathbf{X};\mathbf{h}_k}(\mathbf{h}_k) \mu_{q_{\mathbf{H}};\mathbf{h}_k}(\mathbf{h}_k)}{Z_{b_{\mathbf{h}_k}}} \right\| \right] + c, \end{aligned} \quad (6.21)$$

where $b_{q_{\mathbf{H}_g}}(\mathbf{H}_g)$ denotes the marginalized belief $b_{q_{\mathbf{H}_g}}(\mathbf{H}, \mathbf{X})$ and is calculated by

$$\begin{aligned} b_{q_{\mathbf{H}_g}}(\mathbf{H}_g) &\propto \int b_{q_{\mathbf{H}_g}}(\mathbf{H}, \mathbf{X}) d\mathbf{X} d\mathbf{H}_{\bar{g}} \\ &= p(\mathbf{y}_{p,g}|\mathbf{H}_g) \prod_{k \in G_g} p(\mathbf{h}_k) \mu_{q_{\mathbf{H}},\mathbf{X};\mathbf{h}_k}(\mathbf{h}_k). \end{aligned} \quad (6.22)$$

Since the right argument of (6.21) contains the k -th channel coefficient only, we further marginalize (6.22) to $b_{q_{\mathbf{H}_g}}(\mathbf{h}_k) = \int b_{q_{\mathbf{H}_g}}(\mathbf{H}_g) d\mathbf{h}_{\bar{k}}$. According to the property of KLD, the

Algorithm 6.1 Hybrid EM-EP

Require: $Y_p, Y_d, X_p, \Xi_{h_k}, \sigma_x^2, \sigma_v^2$

- 1: Initialize: $\hat{h}_{k_0:w_{k_0},m_0}, C_{h_{k_0:w_{k_0},m_0}},$
 - 2: $\hat{h}_{k_0:z_{k_0},n_0}, C_{h_{k_0:z_{k_0},n_0}}, \hat{h}_{k_0}, C_{h_{k_0}}$
 - 3: **repeat**
 - 4: [Update based on Pilot Observation $\forall k = 1 : K$]
 - 5: Compute $m_{qH;h_k}$ and $C_{qH;h_k}$ based on (6.23)
 - 6: [Update based on Data Observation $\forall k = 1 : K$]
 - 7: $C_{\tilde{v}_k} = \sigma_v^2 I + \sigma_x^2 \sum_{i \neq k} (C_{qH;h_i} + m_{qH;h_i} m_{qH;h_i}^H)$
 - 8: $\tau_{x_k} = (\hat{h}_k^H C_{\tilde{v}_k}^{-1} \hat{h}_k + \sigma_x^{-2})^{-1}$
 - 9: $\hat{x}_k^T = \tau_{x_k} \hat{h}_k^H C_{\tilde{v}_k}^{-1} Y_d$
 - 10: $C_{qH,X;h_k} = \frac{C_{\tilde{v}_k}}{\hat{x}_k^H \hat{x}_k + T \tau_{x_k}}$
 - 11: $m_{qH,X;h_k} = \frac{Y_d \hat{x}_k^*}{\hat{x}_k^H \hat{x}_k + T \tau_{x_k}}$
 - 12: [Posterior Update $\forall k = 1 : K$]
 - 13: $C_{\hat{h}_k} = (C_{qH;h_k}^{-1} + C_{qH,X}^{-1})^{-1}$
 - 14: $m_{\hat{h}_k} = C_{\hat{h}_k} (C_{qH;h_k}^{-1} m_{qH;h_k} + C_{qH,X}^{-1} m_{qH,X})$
 - 15: **until** Convergence
-

variable level factor is finally updated by $\mu_{qH;h_k}(h_k) = \frac{b_{qH_g}(h_k)}{\mu_{qH,X;h_k}(h_k)} \propto \mathcal{N}(h_k | m_{qH;h_k}, C_{qH;h_k})$ where

$$\begin{aligned}
 C_{qH;h_k} &= \left[\left(\sum_{i \in G_g/\{k\}} (C_{qH,X;h_i}^{-1} + \Xi_{h_i}^{-1})^{-1} + \frac{\sigma_v^2}{\sigma_x^2 P} I \right) + \Xi_{h_k}^{-1} \right]^{-1} \\
 m_{qH;h_k} &= \Xi_{h_k} \left[\Xi_{h_k} + \frac{\sigma_v^2}{\sigma_x^2 P} I + \sum_{i \in G_g/\{k\}} (C_{qH,X;h_i}^{-1} + \Xi_{h_i}^{-1})^{-1} \right]^{-1} \\
 &\quad \cdot \left[\frac{y_{p,g}}{P \sigma_x^2} - \sum_{i \in G_g/\{k\}} \Xi_{h_i} (\Xi_{h_i} + C_{qH,X;h_i})^{-1} m_{qH,X;h_i} \right].
 \end{aligned} \tag{6.23}$$

The approximated belief of h_k is obtained by integrating (6.4),

$$\begin{aligned}
 b(h_k) &= \int b(H, X) d\mathbf{h}_{\bar{k}} dX = \mu_{qH;h_k}(h_k) \mu_{qH,X;h_k}(h_k) \\
 &= \mathcal{N}(h_k | m_{\hat{h}_k}, C_{\hat{h}_k}),
 \end{aligned}$$

where

$$\begin{aligned}
 C_{\hat{h}_k} &= (C_{qH;h_k}^{-1} + C_{qH,X}^{-1})^{-1} \\
 m_{\hat{h}_k} &= C_{\hat{h}_k} (C_{qH;h_k}^{-1} m_{qH;h_k} + C_{qH,X}^{-1} m_{qH,X}).
 \end{aligned} \tag{6.24}$$

A suggested update order is proposed in Algorithm 6.1

6.5 Loop Free Expectation Propagation

The above discussed method is derived from a loopy factorization scheme. To improve the convergence property, we treat H_g and X_g as the atomic variables. The resulting

factorization is obtained by

$$p(\mathbf{Y}_d, \mathbf{y}_{p,g}, \mathbf{H}, \mathbf{X}) = p(\mathbf{Y}_d | \mathbf{H}, \mathbf{X}) \prod_g p(\mathbf{y}_{p,g}, \mathbf{H}_g) p(\mathbf{X}_g).$$

Due to the new factorization, we don't need to approximate $p(\mathbf{y}_{p,g}, \mathbf{H}_g)$ at the level of $\mathbf{h}_k$. We can rewrite (6.2)

$$\mathbf{y}_{p,g} = P\sigma_x^2(\mathbf{1}_{K_g}^T \otimes \mathbf{I}_M)\mathbf{h}_{G_g} + \mathbf{v}_{p,g}, \quad (6.25)$$

where K_g denotes the number of users using the g -th pilot, $\mathbf{1}_{K_g}$ denotes an all-one column vector with K_g entries, and $\mathbf{h}_{G_g} = \text{vec}(\mathbf{H}_g)$. For simplicity, we exploit the notations and define the block prior covariance matrix $\Xi_{G_g} = \text{diag}(\Xi_{\mathbf{h}_{k_1}}, \dots, \Xi_{\mathbf{h}_{k_{K_g}}})$, where $k_1, \dots, k_{K_g}$ are the users using the g -th pilot sequence. We calculate the equivalent prior $p(\mathbf{y}_{p,g}, \mathbf{H}_g) \propto \mathcal{N}(\mathbf{h}_{G_g} | \mathbf{m}_{\mathbf{h}_{G_g}}, \mathbf{C}_{\mathbf{h}_{G_g}})$, where the mean and covariance matrix are obtained by

$$\begin{aligned} \mathbf{C}_{\mathbf{h}_{G_g}} &= \Xi_{\mathbf{h}_{G_g}} - \Xi_{\mathbf{h}_{G_g}}(\mathbf{1}_{K_g} \otimes \mathbf{I}_M) \left[\frac{\sigma_v^2 \mathbf{I}_M}{\sigma_x^2 P} + \sum_{k \in G_g} \Xi_{\mathbf{h}_k} \right]^{-1} \\ &\quad \cdot (\mathbf{1}_{K_g}^T \otimes \mathbf{I}_M) \Xi_{\mathbf{h}_{G_g}} \\ \mathbf{m}_{\mathbf{h}_{G_g}} &= \mathbf{C}_{\mathbf{h}_{G_g}}(\mathbf{1}_{K_g} \otimes \mathbf{I}_M) \frac{\mathbf{y}_{p,g}}{\sigma_v^2}. \end{aligned} \quad (6.26)$$

We will use the equivalent prior $p(\mathbf{y}_{p,g}, \mathbf{H}_g)$ and $p(\mathbf{X}_g)$ to derive an EM algorithm for estimating $\mathbf{H}$.

The observed data symbols can be represented as

$$\mathbf{Y}_d = \mathbf{H}_g \mathbf{X}_g^T + \sum_{j \neq g} \mathbf{H}_j \mathbf{X}_j^T + \mathbf{V}. \quad (6.27)$$

As in the vector EM-EP, we treat $\mathbf{H}_j \mathbf{X}_j^T$ as interference. Due to CLT, we approximate the sum of interference $\sum_{j \neq g} \text{vec}(\mathbf{H}_j \mathbf{X}_j^T)$ as Gaussian. We observe

$$\begin{aligned} \sum_{j \neq g} \text{vec}(\mathbf{H}_j \mathbf{X}_j^T) &= \sum_{i \notin G_g} \text{vec}(\mathbf{h}_i \mathbf{x}_i^T) = \sum_{i \notin G_g} (\mathbf{x}_i \otimes \mathbf{1}_M)(\mathbf{1} \otimes \mathbf{h}_i) \\ &= \sum_{i \notin G_g} (\mathbf{x}_i \otimes \mathbf{h}_i). \end{aligned} \quad (6.28)$$

We define the approximated sum of interference as $\mathbf{Z}_{G_g}$ and its vectorization $\mathbf{z}_{G_g} = \text{vec}(\mathbf{Z}_{G_g})$ such that $\sum_{j \neq g} \text{vec}(\mathbf{H}_j \mathbf{X}_j^T) \simeq \mathbf{z}_{G_g} = \mathcal{N}(\mathbf{z}_{G_g} | \mathbf{0}, \mathbf{C}_{\mathbf{z}_{G_g}})$ where $\mathbf{C}_{\mathbf{z}_{G_g}}$ is the covariance matrix of (6.28),

$$\mathbf{C}_{\mathbf{z}_{G_g}} = \mathbf{I}_T \otimes \left[\sigma_x^2 \sum_{i \notin G_g} (\mathbf{C}_{\mathbf{h}_i} + \mathbf{m}_{\mathbf{h}_i} \mathbf{m}_{\mathbf{h}_i}^H) \right]. \quad (6.29)$$

The noise and interference terms can be combined and interpreted as an equivalent noise. Define $\tilde{\mathbf{V}}_{G_g} = \mathbf{V} + \mathbf{Z}_{G_g}$, and its vectorization $\tilde{\mathbf{v}}_{G_g} = \text{vec}(\tilde{\mathbf{V}}_{G_g})$ can be verified to be a Gaussian $\tilde{\mathbf{v}}_{G_g} \sim \mathcal{N}(\tilde{\mathbf{v}}_{G_g} | \mathbf{0}, \mathbf{I}_T \otimes \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}})$, where

$$\mathbf{C}_{\tilde{\mathbf{v}}_{G_g}} = \sigma_x^2 \sum_{i \notin G_g} (\mathbf{C}_{\mathbf{h}_i} + \mathbf{m}_{\mathbf{h}_i} \mathbf{m}_{\mathbf{h}_i}^H) + \sigma_v^2 \mathbf{I}_M. \quad (6.30)$$

Following the same steps from (6.15) to (6.17) we have

$$-\ln(\hat{p}(\mathbf{Y}_d, \mathbf{y}_{p,g}, \mathbf{H}_g)) \leq \bar{u}_{\mathbf{H}_g}(\mathbf{H}_g | \widehat{\mathbf{H}}_g) - \ln(p(\mathbf{y}_{p,g}, \mathbf{H}_g)) + c, \quad (6.31)$$

where $\widehat{\mathbf{H}}_g$ is the posterior mean of $\mathbf{H}_g$ from previous iteration, and the approximate marginalized pdf $\hat{p}(\mathbf{Y}_d, \mathbf{y}_{p,g}, \mathbf{H}_g) \simeq \int p(\mathbf{Y}_d, \mathbf{y}_{p,g}, \mathbf{H}, \mathbf{X}) d\mathbf{X} d\mathbf{H}_g$ is exact when the CLT approximation $\mathbf{Z}_{G_g}$ is exact, i.e., when $K \rightarrow +\infty$. The term $\bar{u}_{\mathbf{H}_g}(\mathbf{H}_g | \widehat{\mathbf{H}}_g)$ is the effective surrogate function of the negative log-likelihood $-\ln[p(\mathbf{Y}_d | \mathbf{H}_g, \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}})]$, and is defined as

$$\begin{aligned} \bar{u}_{\mathbf{H}_g}(\mathbf{H}_g | \widehat{\mathbf{H}}_g) &= -\mathbb{E}_{\mathbf{X}_g | \mathbf{Y}_d, \widehat{\mathbf{H}}_g, \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}} [\ln(p(\mathbf{Y}_d | \mathbf{X}_g, \mathbf{H}_g, \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}))] \\ &= \|\mathbf{y}_d - (\widehat{\mathbf{X}}_g \otimes \mathbf{I}_M) \mathbf{h}_{G_g}\|_{\mathbf{I}_T \otimes \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1}}^2 + \|\mathbf{h}_{G_g}\|_{T\mathbf{C}_{\widehat{\mathbf{x}}_{G_g}} \otimes \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1}}^2. \end{aligned}$$

where $\widehat{\mathbf{x}}_{G_g}$ and $\mathbf{C}_{\widehat{\mathbf{x}}_{G_g}}$ are the LMMSE results from the model

$$\mathbf{y}_d = (\mathbf{I}_T \otimes \widehat{\mathbf{H}}_g) \mathbf{x}_{G_g} + \tilde{\mathbf{v}}_{G_g}, \quad (6.32)$$

with $\mathbf{x}_{G_g} = \text{vec}(\mathbf{X}_g^T)$ such that $p(\mathbf{X}_g | \mathbf{Y}_d, \widehat{\mathbf{H}}_g, \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}) = \mathcal{N}(\mathbf{x}_{G_g} | \widehat{\mathbf{x}}_{G_g}, \mathbf{I}_T \otimes \mathbf{C}_{\widehat{\mathbf{x}}_{G_g}})$,

$$\begin{aligned} \mathbf{C}_{\widehat{\mathbf{x}}_{G_g}} &= (\widehat{\mathbf{H}}_g^H \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1} \widehat{\mathbf{H}}_g + \sigma_x^{-2} \mathbf{I}_{K_g})^{-1} \\ \widehat{\mathbf{x}}_{G_g} &= [\mathbf{I}_T \otimes (\mathbf{C}_{\widehat{\mathbf{x}}_{G_g}} \widehat{\mathbf{H}}_g^H \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1})] \mathbf{y}_d \end{aligned} \quad (6.33)$$

The matrix mean $\widehat{\mathbf{X}}_g$ is obtained from $\mathbb{E}_{\mathbf{X}_g | \mathbf{Y}_d, \widehat{\mathbf{H}}_g, \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}} [\mathbf{X}_g]$ and can be viewed as a reshaped version of $\widehat{\mathbf{x}}_{G_g}$.

Finally, the surrogate function of the negative log-joint pdf (6.31) is

$$\begin{aligned} -\ln(\hat{p}(\mathbf{Y}_d, \mathbf{y}_{p,g}, \mathbf{H}_g)) &\leq c + \|\mathbf{h}_{G_g} - \mathbf{m}_{\mathbf{h}_{G_g}}\|_{\mathbf{C}_{\mathbf{h}_{G_g}}^{-1}}^2 \\ &+ \|\mathbf{y}_d - (\widehat{\mathbf{X}}_g \otimes \mathbf{I}_M) \mathbf{h}_{G_g}\|_{\mathbf{I}_T \otimes \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1}}^2 + \|\mathbf{h}_{G_g}\|_{T\mathbf{C}_{\widehat{\mathbf{x}}_{G_g}}^* \otimes \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1}}^2. \end{aligned}$$

Optimizing the surrogate function, and we obtain the posterior mean and covariance matrix

$$\begin{aligned} \mathbf{C}_{\widehat{\mathbf{h}}_{G_g}} &= \left[(\widehat{\mathbf{X}}_g^H \widehat{\mathbf{X}}_g + T\mathbf{C}_{\widehat{\mathbf{x}}_{G_g}}^*) \otimes \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1} + \mathbf{C}_{\mathbf{h}_{G_g}}^{-1} \right]^{-1} \\ \mathbf{m}_{\widehat{\mathbf{h}}_{G_g}} &= \mathbf{C}_{\widehat{\mathbf{h}}_{G_g}} \left[(\widehat{\mathbf{X}}_g^H \otimes \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1}) \mathbf{y}_d + \mathbf{C}_{\mathbf{h}_{G_g}}^{-1} \mathbf{m}_{\mathbf{h}_{G_g}} \right]. \end{aligned} \quad (6.34)$$

We propose a possible update order in Algorithm 6.2.

6.6 Simulation Results

In this section, we will verify the algorithm using numerical simulations. We consider a $400 \times 400m^2$ area with $M = 16$ APs and $K = 8$ UTs. The APs are located at the coordinates $(\frac{400}{3}i, \frac{400}{3}j)$, where $i, j \in \{0, \dots, 3\}$. The UTs are uniformly randomly distributed over this area. We use the same expression for the large-scale fading model as in [2],

$$\sigma_{H_{ik}}^2 [\text{dB}] = -30.5 - 36.7 \log_{10}(d_{ik}), \quad (6.35)$$

Algorithm 6.2 MM-algorithm

Require: $\mathbf{Y}_p, \mathbf{Y}_d, \mathbf{X}_p, \mathbf{\Xi}_{\mathbf{h}_k}, \sigma_x^2, \sigma_v^2$

- 1: Initialize: $\hat{\mathbf{h}}_{k_0:\mathbf{w}_{k_0,m_0}}, \mathbf{C}_{\mathbf{h}_{k_0}:\mathbf{w}_{k_0,m_0}},$
 - 2: $\hat{\mathbf{h}}_{k_0:\mathbf{z}_{k_0,n_0}}, \mathbf{C}_{\mathbf{h}_{k_0}:\mathbf{z}_{k_0,n_0}}, \hat{\mathbf{h}}_{k_0}$
 - 3: [Update based on Pilot Observation $\forall g = 1 : P$]
 - 4: Compute $\mathbf{m}_{\mathbf{h}_{G_g}}$ and $\mathbf{C}_{\mathbf{h}_{G_g}}$ based on (6.26)
 - 5: **repeat**
 - 6: [Update based on Data Observation $\forall g = 1 : P$]
 - 7: $\mathbf{C}_{\tilde{\mathbf{v}}_{G_g}} = \sigma_x^2 \sum_{i \notin G_g} (\mathbf{C}_{\mathbf{h}_i} + \mathbf{m}_{\mathbf{h}_i} \mathbf{m}_{\mathbf{h}_i}^H) + \sigma_v^2 \mathbf{I}_M$
 - 8: $\mathbf{C}_{\hat{\mathbf{x}}_{G_g}} = (\hat{\mathbf{H}}_g^H \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1} \hat{\mathbf{H}}_g + \sigma_x^{-2} \mathbf{I}_{K_g})^{-1}$
 - 9: $\hat{\mathbf{x}}_{G_g} = [\mathbf{I}_T \otimes (\mathbf{C}_{\hat{\mathbf{x}}_{G_g}} \hat{\mathbf{H}}_g^H \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1})] \mathbf{y}_d$
 - 10: $\hat{\mathbf{X}}_g^T = \text{vec}_{K_g \times T}^{-1}(\hat{\mathbf{x}}_{G_g})$
 - 11: $\mathbf{C}_{\hat{\mathbf{h}}_{G_g}} = \left[(\hat{\mathbf{X}}_g^H \hat{\mathbf{X}}_g + T \mathbf{C}_{\hat{\mathbf{x}}_{G_g}}^*) \otimes \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1} + \mathbf{C}_{\mathbf{h}_{G_g}}^{-1} \right]^{-1}$
 - 12: $\mathbf{m}_{\hat{\mathbf{h}}_{G_g}} = \mathbf{C}_{\hat{\mathbf{h}}_{G_g}} \left[(\hat{\mathbf{X}}_g^H \otimes \mathbf{C}_{\tilde{\mathbf{v}}_{G_g}}^{-1}) \mathbf{y}_d + \mathbf{C}_{\mathbf{h}_{G_g}}^{-1} \mathbf{m}_{\mathbf{h}_{G_g}} \right]$
 - 13: **until** Convergence
-

where d_{mk} is the distance between AP i and UT k . To induce pilot contamination, the default pilot sequence length is $P = 4$, with the entire transmitted sequence spanning a length of $L = 16$. The four distinct pilot sequences are designed to be mutually orthogonal. Transmission power is set at $\sigma_x^2 = 14(\text{dBm})$ for each UT. We modify the noise power to ensure the $\text{SNR} = \frac{\mathbb{E}[\text{tr}(\mathbf{H}\mathbf{X}_d\mathbf{X}_d^H\mathbf{H}^H)]}{\mathbb{E}[\text{tr}(\mathbf{V}_d\mathbf{V}_d^H)]}$ varies between 10(dB) and 31(dB).

For each SNR level, we maintain consistent positions for all APs and UTs, and conduct simulations across 50 unique scenarios with varying $\mathbf{H}$, $\mathbf{V}_d$, $\mathbf{V}_p$, $\mathbf{V}_d$, and $\mathbf{X}_d$. The metric for evaluating performance is the normalized mean squared error (NMSE) of the channel estimation. It is calculated as $\frac{\sum \text{tr}(\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H)}{\sum \text{tr}(\mathbf{H}\mathbf{H}^H)}$, where $\tilde{\mathbf{H}}$ represents the estimation error, and the summation extends over all 50 distinct realizations. The simulation results are concluded in Fig. 6.1. For comparison, we also plot the results of VL-EP algorithm [20]. The VL-EP algorithm iteratively minimizes the majorizer of $p(\mathbf{h}_{k_0}|\mathbf{p}_{k,m}, \mathbf{r}_{k,n}) \propto \prod_m p(\mathbf{p}_{k_0,m}|\mathbf{h}_{k_0}) \prod_n p(\mathbf{r}_{k_0,n}|\mathbf{h}_{k_0}) p(\mathbf{h}_{k_0})$, functioning similarly to a MAP algorithm. In subsequent iterations, it directly utilizes the posterior of $\mathbf{h}_{k_0}$ to estimate the channel corresponding to other users. However, based on Bayes rule, we should use the partial-posteriors (prior and likelihood from pilot based estimation). We compare these two methods with the normalized approximated CRB [51] which is defined as $\frac{\text{tr}[\mathbf{CRB}]}{\mathbb{E}[\text{tr}(\mathbf{H}\mathbf{H}^H)]}$. The convergence speed of the two algorithms is compared, where we use the default setting and fix the SNR to 20 dB.

6.7 Conclusions

In this chapter, we focus on the joint estimation of the channel and data during uplink transmission in CF MaMIMO systems. To handle the posterior distributions of the per-user received signals correctly, we examine an approximated probabilistic model derived from EP. A hybrid EM EP algorithm aiming to handle the posterior distributions of the per-user received signals correctly is proposed. Whereas, in EM-based algorithm [20] the authors treated the posteriors of different user channels as if they are independent. The effectiveness of the proposed algorithm is demonstrated through numerical simulations.

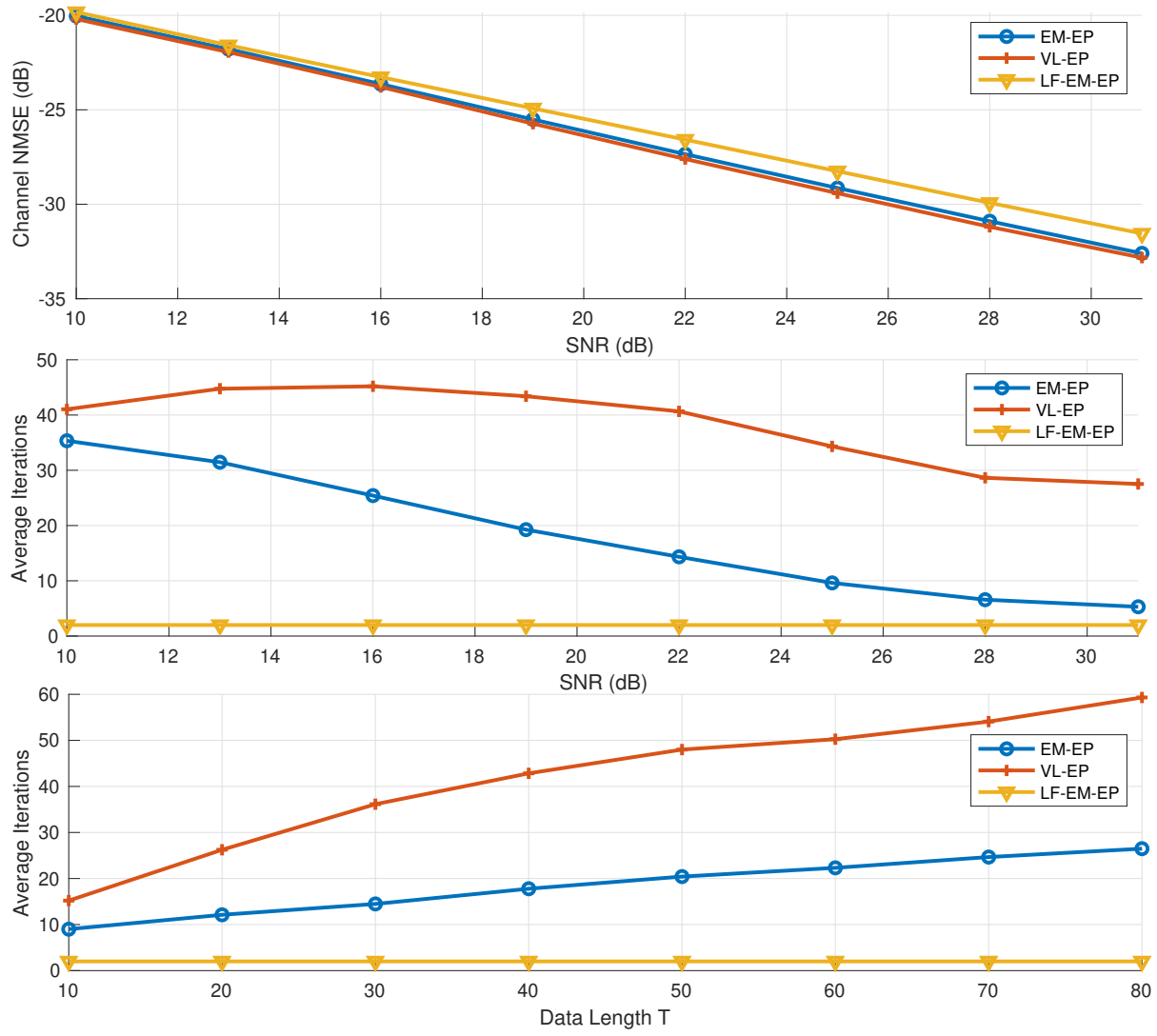


Figure 6.1: NMSE [dB] versus SNR [dB]

Chapter 7

Decentralized Expectation Propagation for Semi-Blind Channel Estimation in Cell-Free Networks

7.1 Introduction

One of the unique features of Cell-Free (CF) Massive MIMO (MaMIMO) networks is user terminals (UTs) in a given area are served by all the access points (APs) in the same area. This leads to the problem of pilot contamination. Addressing this issue, Semi-Blind approaches [50] have been explored to mitigate the effects of pilot contamination. In [50], the authors consider a deterministic approach treating the channel coefficients as deterministic unknown parameters and delve into the analysis of the Cramer-Rao bound and identifiability.

In the context of Bayesian inference, the Semi-Blind approach is modeled as a bilinear inference problem, where the APs must jointly estimate both the channel coefficients and the user signals. Message-passing algorithms play a critical role here. (Loopy) Belief Propagation (BP) [5] is a widely used method in Bayesian inference. It is an iterative method exploiting the structure for a given factorization scheme (e.g., joint/posterior distribution) to lower the computational loads. One of the most potent message-passing algorithms is Expectation Propagation (EP) [9]. Besides exploiting the factorization scheme, EP also projects the beliefs (marginal posterior) to a family of simple distributions. We can consider BP a special case of EP, in which we assume the projection destination set to be the set of all distributions.

A centralized iterative algorithm has been explored based on variable level EP (VL-EP) [20] in which the authors assume Gaussian input and combine VL-EP with Expectation Maximization (EM). A more recent development is the distributed method proposed in [2]. This approach distributes the computational load at the Central Process Unit (CPU) by enabling each Access Point (AP) to carry out part of the computation.

In [27], the author introduces Decentralized Generalized Approximate Message Passing (D-GAMP). This method is a hybrid of Consensus Propagation [26] and Approximate Message Passing (AMP), effectively eliminating the need for CPU.

7.1.1 Main Contributions

This chapter presents a simplified, decentralized, EP-based method designed to address the Semi-Blind estimation problem in communication systems. By utilizing orthogonal pilots, we are able to decouple the channels for different users into mutually exclusive groups, which reduces computational complexity. To further decrease computational demands, we integrate Expectation Propagation (EP) with Central Limit Theory (CLT), treating the interference as noise. Drawing inspiration from [23], we introduce further simplifications through scale analysis. Additionally, to lessen the load on the central processing unit (CPU), we explore a decentralized scheme.

7.2 System Model

We consider a semi-blind signal model containing L APs. At the l -th AP,

$$[\mathbf{Y}_{p,l} \ \mathbf{Y}_l] = \mathbf{H}_l [\mathbf{X}_p \ \mathbf{X}] + [\mathbf{V}_{p,l} \ \mathbf{V}_l]. \quad (7.1)$$

The received signals are composed of pilot part $\mathbf{Y}_{p,l} \in \mathbb{C}^{N \times P}$ and data part $\mathbf{Y}_l \in \mathbb{C}^{N \times T}$. The channels between different users are considered independent Gaussian i.e. $\text{vec}(\mathbf{H}_l) \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Xi}_l)$ where $\mathbf{\Xi}_l \in \mathbb{C}^{NK \times NK}$ is a block diagonal matrix of K blocks $\mathbf{\Xi}_{h_{lk}} \in \mathbb{C}^{N \times N}$. The transmitted symbols can be decomposed as pilot symbols $\mathbf{X}_p \in \mathcal{S}^{K \times P}$ and data symbols $\mathbf{X} \in \mathcal{S}^{K \times T}$, where $\mathcal{S}$ is the constellation set. We assume that the elements x_{kt} in $\mathbf{X}$ follow the categorical distribution $p(x_{kt})$. The signal power is denoted as σ_x^2 . The noise is considered as i.i.d. Gaussian distribution, and thus, $\text{vec}([\mathbf{V}_{p,l} \ \mathbf{V}_l]) \sim \mathcal{CN}(\mathbf{0}, \sigma_v^2 \mathbf{I})$.

7.2.1 Orthogonal Pilot sequences

If orthogonal pilot sequences are used, we can first preprocess the pilot observation by right multiplying it with $\mathbf{x}_{p,g}^*$ which is the conjugated g -th pilot sequence. This results in an equivalent observation $\mathbf{y}_{p,lg}$

$$\mathbf{y}_{p,lg} = \mathbf{Y}_{p,l} \mathbf{x}_{p,g}^* = \sum_{k \in G_g} P \sigma_x^2 \mathbf{h}_{lk} + \mathbf{v}_{p,lg} \quad (7.2)$$

where $\mathbf{v}_{p,lg} = \mathbf{V}_{p,l} \mathbf{x}_{p,g}^* \sim \mathcal{N}(\mathbf{v}_{p,lg} | \mathbf{0}, P \sigma_x^2 \sigma_v^2 \mathbf{I})$, G_g denote the set of users using the g -th pilot sequence. We observe that every $\mathbf{h}_{lk}$ occurs only in one group G_g , and the cross-correlation $\mathbb{E}[\mathbf{v}_{p,lg} \mathbf{v}_{p,lg'}^H]$ is an all-zero matrix for all $g \neq g'$. Therefore, the observations $\mathbf{y}_{p,lg}$ and $\mathbf{y}_{p,lg'}$ are independent. With orthogonal pilots, the factorization scheme is derived as

$$\begin{aligned} & p(\{\mathbf{y}_{p,lg}\}, \{\mathbf{Y}_l\}, \{\mathbf{H}_l\}, \mathbf{X}, \{\mathbf{V}_l\}) \\ &= \prod_{k,t} p(x_{kt}) \prod_l \prod_{t_1=1}^T p(\mathbf{y}_{lt_1} | \mathbf{H}_l, \mathbf{x}_{:t_1}) \prod_g p(\mathbf{y}_{p,lg}, \mathbf{H}_{lg}) \end{aligned} \quad (7.3)$$

where $\mathbf{H}_{lg}$ is a matrix collecting all $k \in G_g, \mathbf{h}_{lk}$ as its column vectors, and $\mathbf{x}_{:t_1}$ denotes the t_1 -th column of $\mathbf{X}$. We will base our EP (BP) algorithm based on this factorization scheme.

7.2.2 Expectation Propagation on Semi-Blind structure

For simplicity, we denote the factors in the factorization scheme (7.3) as

$$\Psi_{1,kt} = p(x_{kt}); \ \Psi_{2,lt} = p(\mathbf{y}_{lt} | \mathbf{H}_l, \mathbf{x}_{:t}); \ \Psi_{3,lg} = p(\mathbf{y}_{p,lg}, \mathbf{H}_{lg}).$$

The factor graph for (7.3) is illustrated in Fig. 7.1.

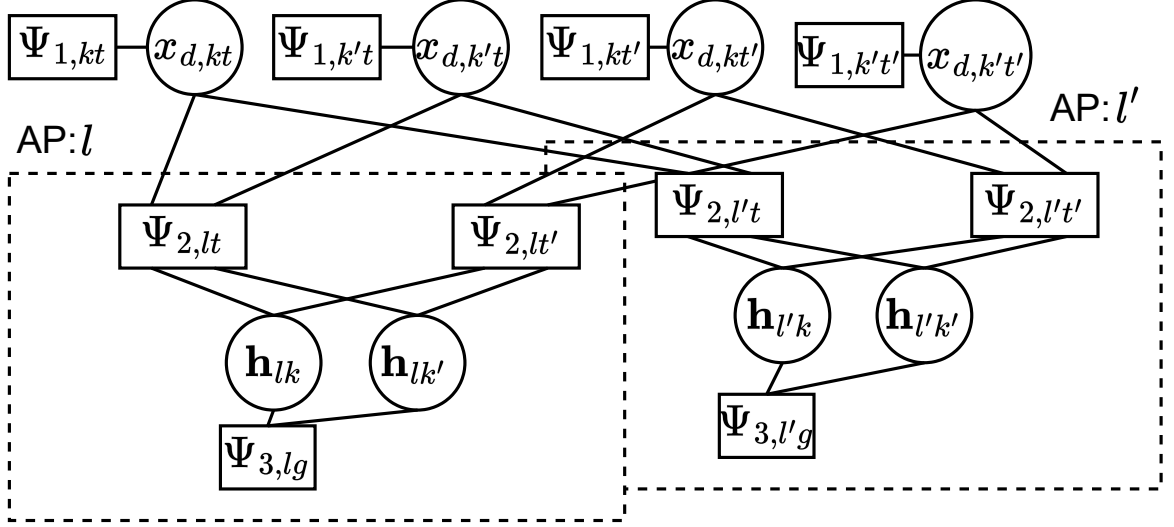


Figure 7.1: Partial factor graph

7.3 Message Passing Derivations

In this chapter, we use EP to estimate channel coefficients $\mathbf{h}_{lk}$ and BP to estimate the data symbols x_{kt} . Furthermore, we specify the projection family of EP in this chapter to be Gaussian distributions with diagonal covariance matrices. Now, we will examine each factor and derive its outbound message.

The message from $\Psi_{1,kt}$ to x_{kt} can be computed directly since no projection is needed, i.e., $\mu_{\Psi_{1,kt};x_{kt}}(x_{kt}) = p(x_{kt})$.

7.3.1 Message from $\Psi_{2,lt}$ to x_{kt}

Following the EP rule for computing the feedback message, we first need to compute the belief. The extrinsic node $\Psi_{2,lt}$, which is updated by

$$\begin{aligned} \mu_{x_{kt};\Psi_{2,lt}}(x_{kt}) &\propto p(x_{kt}) \prod_{l' \neq l} \mu_{\Psi_{2,l't};x_{kt}}(x_{kt}) \\ \mu_{\mathbf{h}_{lk};\Psi_{2,lt}}(\mathbf{h}_{lk}) &\propto \mu_{\Psi_{3,lg};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) \prod_{t' \neq t} \mu_{\Psi_{2,l't'};\mathbf{h}_{lk}}(\mathbf{h}_{lk}), \end{aligned} \quad (7.4)$$

where the extrinsic of $\mathbf{h}_{lk}$ is Gaussian $\mu_{\mathbf{h}_{lk};\Psi_{2,lt}}(\mathbf{h}_{lk}) = \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}, \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}})$ with

$$\begin{aligned} \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} &= \left(\mathbf{C}_{\Psi_{3,lg};\mathbf{h}_{lk}}^{-1} + \sum_{t' \neq t} \mathbf{C}_{\Psi_{2,l't'};\mathbf{h}_{lk}}^{-1} \right)^{-1} \\ \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} &= \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} \left(\mathbf{C}_{\Psi_{3,lg};\mathbf{h}_{lk}}^{-1} \mathbf{m}_{\Psi_{3,lg};\mathbf{h}_{lk}} \right. \\ &\quad \left. + \sum_{t' \neq t} \mathbf{C}_{\Psi_{2,l't'};\mathbf{h}_{lk}}^{-1} \mathbf{m}_{\Psi_{2,l't'};\mathbf{h}_{lk}} \right) \end{aligned}$$

According to the EP rule, the message from $\Psi_{2,lt}$ to x_{kt} is

$$\mu_{\Psi_{2,lt};x_{kt}}(x_{kt}) \propto \frac{\text{proj}[b_{\Psi_{2,lt};x_{kt}}(x_{kt})]}{\mu_{x_{kt};\Psi_{2,lt}}(x_{kt})}, \quad (7.5)$$

where the belief (approximated posterior) at $\Psi_{2,lt}$ is defined as $b_{\Psi_{2,lt};x_{kt}}(x_{kt})$ with

$$\begin{aligned}
& b_{\Psi_{2,lt};x_{kt}}(x_{kt}) \mu_{x_{kt};\Psi_{2,lt}}(x_{kt}) \sum_{\mathbf{x}_{\bar{k}t}} \int p(\mathbf{y}_{lt} | x_{kt} \mathbf{h}_{lk} + \sum_{i \neq k} x_{it} \mathbf{h}_{li}) \\
& \cdot \mu_{\mathbf{h}_{lk};\Psi_{2,lt}}(\mathbf{h}_{lk}) \prod_{i \neq k} \mu_{\mathbf{h}_{li};\Psi_{2,lt}}(\mathbf{h}_{li}) \mu_{x_{it};\Psi_{2,lt}}(x_{it}) d\mathbf{H}_l.
\end{aligned} \tag{7.6}$$

We use the notation $\mathbf{x}_{\bar{k}t}$ to denote all the elements in $\mathbf{x}_{:t}$ except the k -th element. The integral (and summation) in (7.6) can be considered as a marginalization operation. Furthermore, we can view the extrinsic messages as hypothetical priors. Due to CLT, we approximate $\sum_{i \neq k} x_{it} \mathbf{h}_{li}$ to a Gaussian where $x_{it} \sim \mu_{x_{it};\Psi_{2,lt}}(x_{it})$, $\mathbf{h}_{li} \sim \mu_{\mathbf{h}_{li};\Psi_{2,lt}}(\mathbf{h}_{li})$ [15]. Therefore, (7.6) becomes

$$\begin{aligned}
& b_{\Psi_{2,lt};x_{kt}}(x_{kt}) \propto \mu_{x_{kt};\Psi_{2,lt}}(x_{kt}) \\
& \cdot \iint p(\mathbf{y}_{lt} | x_{kt} \mathbf{h}_{lk} + \mathbf{z}_{lkt}) \mu_{\mathbf{z}_{lkt}}(\mathbf{z}_{lkt}) d\mathbf{z}_{lkt} \cdot \mu_{\mathbf{h}_{lk};\Psi_{2,lt}}(\mathbf{h}_{lk}) d\mathbf{h}_{lk},
\end{aligned} \tag{7.7}$$

where $\mu_{\mathbf{z}_{lkt}}(\mathbf{z}_{lkt}) = \mathcal{CN}(\mathbf{z}_{lkt} | \mathbf{m}_{\mathbf{z}_{lkt}}, \mathbf{C}_{\mathbf{z}_{lkt}})$ with

$$\begin{aligned}
\mathbf{m}_{\mathbf{z}_{lkt}} &= \sum_{i \neq k} m_{x_{it};\Psi_{2,lt}} \mathbf{m}_{\mathbf{h}_{li};\Psi_{2,lt}} \\
\mathbf{C}_{\mathbf{z}_{lkt}} &= \sum_{i \neq k} r_{x_{it};\Psi_{2,lt}} \mathbf{C}_{\mathbf{h}_{li};\Psi_{2,lt}} + \tau_{x_{it};\Psi_{2,lt}} \mathbf{m}_{\mathbf{h}_{li};\Psi_{2,lt}} \mathbf{m}_{\mathbf{h}_{li};\Psi_{2,lt}}^H
\end{aligned} \tag{7.8}$$

where $m_{x_{it};\Psi_{2,lt}}$, $\tau_{x_{it};\Psi_{2,lt}}$ and $r_{x_{it};\Psi_{2,lt}}$ are the mean, variance and second-order moment of the normalized message $\mu_{x_{it};\Psi_{2,lt}}$. By applying the Gaussian reproduction lemma [15] and the fact that Gaussian distribution integrates to one, the belief (7.7) becomes

$$\begin{aligned}
& b_{\Psi_{2,lt};x_{kt}}(x_{kt}) \propto \mathcal{CN}(\mathbf{0} | \mathbf{y}_{lt} - \mathbf{m}_{\mathbf{z}_{lkt}} - x_{kt} \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}, \\
& \mathbf{C}_v + \mathbf{C}_{\mathbf{z}_{lkt}} + |x_{kt}|^2 \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}) \cdot \mu_{x_{kt};\Psi_{2,lt}}(x_{kt}).
\end{aligned} \tag{7.9}$$

Therefore, by BP rules, the outbound message is

$$\begin{aligned}
& \mu_{\Psi_{2,lt};x_{kt}}(x_{kt}) \propto \mathcal{CN}(\mathbf{0} | \mathbf{y}_{lt} - \mathbf{m}_{\mathbf{z}_{lkt}} - x_{kt} \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}, \\
& \mathbf{C}_v + \mathbf{C}_{\mathbf{z}_{lkt}} + |x_{kt}|^2 \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}})
\end{aligned} \tag{7.10}$$

7.3.2 Message from $\Psi_{2,lt}$ to $\mathbf{h}_{lk}$

Based on EP feedback message rules, the message to $\mathbf{h}_{lk}$ is

$$\mu_{\Psi_{2,lt};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) \propto \frac{\text{proj}[b_{\Psi_{2,lt};\mathbf{h}_{lk}}(\mathbf{h}_{lk})]}{\mu_{\mathbf{h}_{lk};\Psi_{2,lt}}(\mathbf{h}_{lk})}, \tag{7.11}$$

where the belief is defined as

$$\begin{aligned}
& b_{\Psi_{2,lt};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) \propto \sum_{\mathbf{x}_{:t}} \int p(\mathbf{y}_{lt} | \sum_i x_{it} \mathbf{h}_{li}) \\
& \cdot \prod_i \mu_{\mathbf{h}_{li};\Psi_{2,lt}}(\mathbf{h}_{li}) \mu_{x_{it};\Psi_{2,lt}}(x_{it}) d\mathbf{h}_{\bar{l}k}
\end{aligned} \tag{7.12}$$

We use $\mathbf{h}_{\bar{l}k}$ to denote all the column vectors in $\mathbf{H}_l$ except the k -th column. By using the same approach from (7.6) to (7.9), and separating the terms that contains only x_{kt} [15] [2], the belief (7.12) becomes

$$\begin{aligned}
& b_{\Psi_{2,lt};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) \\
& = \mathbb{E}_{b_{\Psi_{2,lt};x_{kt}}} \{ \mathcal{CN}[\mathbf{h}_{lk} | \mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x_{kt}), \mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x_{kt})] \}
\end{aligned} \tag{7.13}$$

where $\mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(\cdot)$ and $\mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(\cdot)$ are defined as

$$\begin{aligned}\mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x) &= [|x|^2(\mathbf{C}_v + \mathbf{C}_{\mathbf{z}_{lkt}})^{-1} + \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}^{-1}]^{-1} \\ \mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x) &= \mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x) \left[\mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}^{-1} \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} \right. \\ &\quad \left. + |x|^2(\mathbf{C}_v + \mathbf{C}_{\mathbf{z}_{lkt}})^{-1} \frac{\mathbf{y}_{lt} - \mathbf{m}_{\mathbf{z}_{lkt}}}{x} \right],\end{aligned}\tag{7.14}$$

where $\mathbf{C}_v = \sigma_v^2 \mathbf{I}$. The mean $\mathbf{m}_{\hat{\mathbf{h}}_{lk}^2}$ and covariance $\mathbf{C}_{\hat{\mathbf{h}}_{lk}^2}$ of the belief distribution (7.13) are

$$\begin{aligned}\mathbf{m}_{\hat{\mathbf{h}}_{lk}^2} &= \mathbb{E}_{b_{\Psi_{2,lt};x_{kt}}}[\mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x_{kt})] \\ \mathbf{C}'_{\hat{\mathbf{h}}_{lk}^2} &= \mathbb{E}_{b_{\Psi_{2,lt};x_{kt}}}[\mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x_{kt}) \\ &\quad + \mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x_{kt})\mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x_{kt})^H] - \mathbf{m}_{\hat{\mathbf{h}}_{lk}^2} \mathbf{m}_{\hat{\mathbf{h}}_{lk}^2}^H.\end{aligned}\tag{7.15}$$

Project the belief at $\Psi_{2,lt}$ to Gaussian with diagonal covariance matrix $\text{proj}[b_{\Psi_{2,lt};\mathbf{h}_{lk}}(\mathbf{h}_{lk})] = \mathcal{CN}(\mathbf{h}_{lk}|\mathbf{m}_{\hat{\mathbf{h}}_{lk}^2}, \mathbf{C}_{\hat{\mathbf{h}}_{lk}^2})$, where $\mathbf{C}_{\hat{\mathbf{h}}_{lk}^2}$ is a diagonal matrix with the same diagonal elements as $\mathbf{C}'_{\hat{\mathbf{h}}_{lk}^2}$. Finally, the message from $\Psi_{2,lt}$ to $\mathbf{h}_{lk}$ is

$$\begin{aligned}\mu_{\Psi_{2,lt};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) &= \mathcal{CN}(\mathbf{h}_{lk}|\mathbf{m}_{\Psi_{2,lt};\mathbf{h}_{lk}}, \mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}}) \\ &\propto \frac{\mathcal{CN}(\mathbf{h}_{lk}|\mathbf{m}_{\hat{\mathbf{h}}_{lk}^2}, \mathbf{C}_{\hat{\mathbf{h}}_{lk}^2})}{\mathcal{CN}(\mathbf{h}_{lk}|\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}, \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}})}.\end{aligned}\tag{7.16}$$

7.3.3 Message form $\Psi_{3,lg}$ to $\mathbf{h}_{lk}$

We assume $k \in G_g$. The extrinsic at $\Psi_{3,lg}$ is updated by

$$\mu_{\mathbf{h}_{lk};\Psi_{3,lg}}(\mathbf{h}_{lk}) \propto \prod_t \mu_{\Psi_{2,lt};\mathbf{h}_{lk}}(\mathbf{h}_{lk}).\tag{7.17}$$

We denote this extrinsic message as $\mu_{\mathbf{h}_{lk};\Psi_{3,lg}}(\mathbf{h}_{lk}) = \mathcal{CN}(\mathbf{h}_{lk}|\mathbf{m}_{\mathbf{h}_{lk};\Psi_{3,lg}}, \mathbf{C}_{\mathbf{h}_{lk};\Psi_{3,lg}})$ with

$$\begin{aligned}\mathbf{C}_{\mathbf{h}_{lk};\Psi_{3,lg}} &= \left(\sum_t \mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}}^{-1} \right)^{-1} \\ \mathbf{m}_{\mathbf{h}_{lk};\Psi_{3,lg}} &= \mathbf{C}_{\mathbf{h}_{lk};\Psi_{3,lg}} \left(\sum_t \mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}}^{-1} \mathbf{m}_{\Psi_{2,lt};\mathbf{h}_{lk}} \right).\end{aligned}$$

The belief of $\mathbf{h}_{lg}$ at the $\Psi_{3,lg}$ is

$$b_{\Psi_{3,lg}}(\mathbf{h}_{lg}) \propto p(\mathbf{y}_{p,lg}, \mathbf{h}_{lg}) \prod_{k \in G_g} p(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk};\Psi_{3,lg}}(\mathbf{h}_{lk}).\tag{7.18}$$

All the factors appearing in (7.18) are Gaussian pdfs with diagonal covariance matrices. Therefore, the projection of $b_{\Psi_{3,lg}}(\mathbf{h}_{lk})$ results to itself. For simplicity, we define a hypothetical prior $q_{\mathbf{h}_{lk}|\mathbf{Y}_d}$ for $\mathbf{h}_{lk}$ in (7.18) as

$$\begin{aligned}q_{\mathbf{h}_{lk}|\mathbf{Y}_d}(\mathbf{h}_{lk}) &= \mathcal{N}(\mathbf{h}_{lk}|\mathbf{m}_{\mathbf{h}_{lk}|\mathbf{Y}_d}, \mathbf{C}_{\mathbf{h}_{lk}|\mathbf{Y}_d}) \\ &\propto p(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk};\Psi_{3,lg}}(\mathbf{h}_{lk}),\end{aligned}\tag{7.19}$$

where

$$\begin{aligned}\mathbf{C}_{\mathbf{h}_{lk}|\mathbf{Y}_d} &= (\mathbf{\Xi}_{\mathbf{h}_{lk}}^{-1} + \mathbf{C}_{\mathbf{h}_{lk};\Psi_{3,lg}}^{-1})^{-1} \\ \mathbf{m}_{\mathbf{h}_{lk}|\mathbf{Y}_d} &= \mathbf{C}_{\mathbf{h}_{lk}|\mathbf{Y}_d} \mathbf{C}_{\mathbf{h}_{lk};\Psi_{3,lg}}^{-1} \mathbf{m}_{\mathbf{h}_{lk};\Psi_{3,lg}}\end{aligned}\tag{7.20}$$

The message from factor node $\Psi_{3,lg}$ to $\mathbf{h}_{lk}$ can be derived as

$$\begin{aligned}\mu_{\Psi_{3,lg};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) &\propto \frac{\int b_{\Psi_{3,lg}}(\mathbf{h}_{lg}) d\mathbf{h}_{l\bar{k}}}{\mu_{\mathbf{h}_{lk};\Psi_{3,lg}}(\mathbf{h}_{lk})} \\ &\propto p(\mathbf{h}_{lg}) \frac{\prod_{k \in G_g} q_{\mathbf{h}_{lk}|\mathbf{Y}_d}(\mathbf{h}_{lk}) d\mathbf{h}_{l\bar{k}}}{q_{\mathbf{h}_{lk}|\mathbf{Y}_d}(\mathbf{h}_{lk})}\end{aligned}\quad (7.21)$$

The fraction operation in the second line of (7.21) can be interpreted as component-wise conditionally-unbiased LMMSE estimation [34]. Therefore, the message from $\Psi_{3,lg}$ to $\mathbf{h}_{lk}$ is

$$\mu_{\Psi_{3,lg};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) = \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{\Psi_{3,lg};\mathbf{h}_{lk}}, \mathbf{C}_{\Psi_{3,lg};\mathbf{h}_{lk}}), \quad (7.22)$$

where

$$\begin{aligned}\mathbf{C}_{\Psi_{3,lg};\mathbf{h}_{lk}} &= \left[\mathbf{\Xi}_{\mathbf{h}_{lk}}^{-1} + \left(\frac{\sigma_v^2}{\sigma_x^2 P} \mathbf{I} + \sum_{k' \in G_g/\{k\}} \mathbf{C}_{\mathbf{h}_{lk'}|\mathbf{Y}_d} \right)^{-1} \right]^{-1} \\ \mathbf{m}_{\Psi_{3,lg};\mathbf{h}_{lk}} &= \mathbf{\Xi}_{\mathbf{h}_{lk}} \left(\frac{\sigma_v^2}{\sigma_x^2 P} \mathbf{I} + \sum_{k' \in G_g/\{k\}} \mathbf{C}_{\mathbf{h}_{lk'}|\mathbf{Y}_d} + \mathbf{\Xi}_{\mathbf{h}_{lk}} \right)^{-1} \\ &\quad \cdot \left(\frac{1}{\sigma_x^2 P} \mathbf{y}_{p,lg} - \sum_{k' \in G_g/\{k\}} \mathbf{m}_{\mathbf{h}_{lk'}|\mathbf{Y}_d} \right).\end{aligned}\quad (7.23)$$

7.4 Asymptotic Behaviors in Large Systems

For scalable systems, $L \rightarrow \infty$ while $K = c_1 L$, $P = c_2 L$, $T = c_3 L$, where c_1, c_2, c_3 are some positive constants, we assume the channel coefficients $\forall l, n, k, \mathbb{E}[|h_{lnk}|^2] = O(\frac{1}{L})$ data constellation symbols $\forall s \in \mathcal{S}, s = O(1), 1/s = O(1)$. Furthermore, we assume the noise power scales as $\sigma_v^2 = O(1)$, $\sigma_v^{-2} = O(1)$. For simplicity, we define big-O-notations with matrix parameters to represent the element-wise asymptotic behavior, i.e., for matrices $\mathbf{A}, \mathbf{B}$ of the same size, we have $\mathbf{A} = O(\mathbf{B}) \Leftrightarrow \forall i, j, [\mathbf{A}]_{ij} = O([\mathbf{B}]_{ij})$.

Assumption 1. We assume that $\mathbb{E}(\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}) = \mathbf{0}$, $\mathbb{E}[m_{x_{it};\Psi_{2,lt}} \mathbf{m}_{\mathbf{h}_{li};\Psi_{2,lt}}] = \mathbf{0}$, and $\forall i \neq j, \mathbb{E}([\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}^H]_{ij}) = 0$.

Property 1. For invertible matrices $\mathbf{A}, \mathbf{B}$, we have $(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1} = \mathbf{A}(\mathbf{A} + \mathbf{B})^{-1} \mathbf{B} = \mathbf{B}(\mathbf{A} + \mathbf{B})^{-1} \mathbf{A}$.

Lemma 4. With proper initialization and Assumption 1, in each iteration, the updates satisfy $\mathbf{m}_{\mathbf{z},lkt} = O(1)$, $\mathbf{C}_{\mathbf{z},lkt} = O(\mathbf{I}) + O(\frac{1 \cdot 1^H}{\sqrt{L}})$, $\mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x) = O(\frac{1}{\sqrt{L}})$, $\mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x) = O(\frac{\mathbf{I}}{L}) + O(\frac{1 \cdot 1^H}{L^2})$, $\mathbf{m}_{\Psi_{2,lt};\mathbf{h}_{lk}} = O(\frac{1}{\sqrt{L}})$, $\mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}} = O(\mathbf{I})$, $\mathbf{m}_{\Psi_{3,lg};\mathbf{h}_{lk}} = O(\frac{1}{\sqrt{L}})$, $\mathbf{C}_{\Psi_{3,lg};\mathbf{h}_{lk}} = O(\frac{\mathbf{I}}{L})$, $\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} = O(\frac{1}{\sqrt{L}})$, $\mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} = O(\frac{\mathbf{I}}{L})$. Furthermore, $\mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}}^{-1} = O(\mathbf{I})$.

Proof. We prove this lemma by mathematical induction. Due to a proper initialization, we can assume the messages $\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}$, $\mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}$, $\mathbf{m}_{\Psi_{2,lt};\mathbf{h}_{lk}}$, $\mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}}$, $\mathbf{m}_{\Psi_{3,lg};\mathbf{h}_{lk}}$, $\mathbf{C}_{\Psi_{3,lg};\mathbf{h}_{lk}}$ are initialized with the above-mentioned scales.

Then, we assume the lemma holds for the previous iterations and investigate the updates in the next iteration.

We first look at the update of $\mathbf{m}_{\mathbf{z},lkt}$, $\mathbf{C}_{\mathbf{z},lkt}$, which are updated according to (7.8). Similar to [13], we assume that $\forall i, m_{x_{it};\Psi_{2,lt}}$ are weakly independent of $\mathbf{m}_{\mathbf{h}_{li};\Psi_{2,lt}}$. Since the

elements in the constellation set scale with $O(1)$, we know $m_{x_{it};\Psi_{2,lt}} = O(1)$. According to induction assumptions, the extrinsic mean $\mathbf{m}_{\mathbf{h}_{li};\Psi_{2,lt}} = O(\frac{1}{\sqrt{L}})$. Due to Assumption 1, we use the results from [13, Lemma 1] to obtain $\mathbf{m}_{\mathbf{z}_{lkt}} = \sum_{i \neq k} m_{x_{it};\Psi_{2,lt}} \mathbf{m}_{\mathbf{h}_{li};\Psi_{2,lt}} = O(1)$ and the covariance matrix $\mathbf{C}_{\mathbf{z}_{lkt}} = O(\mathbf{I}) + O(\frac{1}{\sqrt{L}})$. For simplicity, we denote the diagonal terms of $\mathbf{C}_{\mathbf{z}_{lkt}}$ in (7.8) as $\mathbf{D}_{\mathbf{z}_{lkt}} = \sum_{i \neq k} r_{x_{it};\Psi_{2,lt}} \mathbf{C}_{\mathbf{h}_{li};\Psi_{2,lt}}$, and denote $\mathbf{B}_{\mathbf{z}_{lkt}} = \sum_{i \neq k} \tau_{x_{it};\Psi_{2,lt}} \mathbf{m}_{\mathbf{h}_{li};\Psi_{2,lt}} \mathbf{m}_{\mathbf{h}_{li};\Psi_{2,lt}}^H = O(\mathbf{I}) + O(\frac{1}{\sqrt{L}})$. Thus, with these notations, $\mathbf{C}_{\mathbf{z}_{lkt}} = \mathbf{D}_{\mathbf{z}_{lkt}} + \mathbf{B}_{\mathbf{z}_{lkt}}$.

Now we investigate the update of $\mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x)$, $\mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x)$ in (7.14). By matrix inversion lemma,

$$\begin{aligned} \mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x) &= \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} \\ &\quad - \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} \mathbf{Q}_{lkt}^{-\frac{H}{2}} \mathbf{W}_{lkt} (\mathbf{\Lambda}_{lkt} + \mathbf{I})^{-1} \mathbf{W}_{lkt}^H \mathbf{Q}_{lkt}^{-\frac{1}{2}} \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}, \end{aligned} \quad (7.24)$$

where we define the positive semi-definite diagonal matrix $\mathbf{Q}_{lkt} = \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} + \frac{1}{|x|^2} (\mathbf{C}_{\mathbf{v}} + \mathbf{D}_{\mathbf{z}_{lkt}}) = O(\mathbf{I})$ and $\mathbf{Q}_{lkt}^{\frac{1}{2}} \mathbf{Q}_{lkt}^{\frac{H}{2}} = \mathbf{Q}_{lkt}$. By eigendecomposition, we define $\frac{1}{|x|^2} \mathbf{Q}_{lkt}^{-\frac{1}{2}} \mathbf{B}_{\mathbf{z}_{lkt}} \mathbf{Q}_{lkt}^{-\frac{H}{2}} = \mathbf{W}_{lkt} \mathbf{\Lambda}_{lkt} \mathbf{W}_{lkt}^H$, $\mathbf{I} = \mathbf{W}_{lkt} \mathbf{W}_{lkt}^H$. Therefore, $\mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x) = O(\frac{\mathbf{I}}{L}) + O(\frac{1}{L^2})$. We find the update of $\mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}}$ in (7.14) is dominated by the first term $\mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x) \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}^{-1} \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}$. By neglecting higher order infinitesimal terms, we have

$$\begin{aligned} \mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}} &\simeq \mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x) \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}^{-1} \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} = [\mathbf{I} + \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}^{\frac{1}{2}} \\ &\quad \cdot (\mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}^{\frac{H}{2}} \mathbf{C}_{\text{in},lkt}^{-1} \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}^{\frac{1}{2}} + \mathbf{I})^{-1} \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}^{-\frac{H}{2}} \mathbf{C}_{\text{in},lkt}^{-1}] \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} \\ &\simeq \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} = O(\frac{1}{\sqrt{L}}). \end{aligned}$$

To study the messages $\mathbf{m}_{\Psi_{2,lt};\mathbf{h}_{lk}}$, $\mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}}$, we first investigate the approximated (projected) belief of $\mathbf{h}_{lk}$ at $\Psi_{2,lt}$. From the previous discussion, $\mathbf{C}'_{\hat{\mathbf{h}}_{lk}^2} \simeq \mathbb{E}_{b_{\Psi_{2,lt};x_{kt}}} [\mathbf{C}_{\hat{\mathbf{h}}_{lk}|x_{kt}}(x_{kt})]$, and thus,

$$\mathbf{C}'_{\hat{\mathbf{h}}_{lk}^2} \simeq \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} - \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} \cdot \mathbf{F} \cdot \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}, \quad (7.25)$$

where $\mathbf{F} = \sum_{x \in \mathcal{S}} b_{\Psi_{2,lt};x_{kt}}(x) \left[\frac{1}{|x|^2} \mathbf{B}_{\mathbf{z}_{lkt}} + \mathbf{Q}_{lkt} \right]^{-1}$. Thanks to the projection in EP, we are only interested in the diagonal elements of $\mathbf{C}'_{\hat{\mathbf{h}}_{lk}^2}$. With the approximation $\mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}} \simeq \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}$, the n -th diagonal term reads

$$\begin{aligned} [\mathbf{C}'_{\hat{\mathbf{h}}_{lk}^2}]_{nn} &= [\mathbf{C}_{\hat{\mathbf{h}}_{lk}^2}]_{nn} \simeq \tau_{h_{lnk};\Psi_{2,lt}} - \tau_{h_{lnk};\Psi_{2,lt}}^2 [\mathbf{F}]_{nn} \\ &= [([\mathbf{F}]_{nn}^{-1} - \tau_{h_{lnk};\Psi_{2,lt}})^{-1} + \tau_{h_{lnk};\Psi_{2,lt}}^{-1}]^{-1} \end{aligned} \quad (7.26)$$

From the same analysis in (7.24), we know $\mathbf{Q}_{lkt}$ is asymptotic upper and lower bounded. Since $\mathbf{B}_{\mathbf{z}_{lkt}}$ is positive semi-definite, we have $[\mathbf{F}]_{nn} = O(1)$ and $[\mathbf{F}]_{nn}^{-1} = O(1)$. Substitute (7.26) into (7.16), and we obtain $[\mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}}]_{nn} \simeq [\mathbf{F}]_{nn}^{-1} - \tau_{h_{lnk};\Psi_{2,lt}}$. Thus, $\mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}} = O(\mathbf{I})$ and $\mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}}^{-1} = O(\mathbf{I})$. Since $\mathbf{m}_{\hat{\mathbf{h}}_{lk}|x_{kt}} \simeq \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}$, it is straightforward to see $\mathbf{m}_{\Psi_{2,lt};\mathbf{h}_{lk}} = O(\frac{1}{\sqrt{L}})$.

The message covariance matrices $\mathbf{C}_{\Psi_{3,lg};\mathbf{h}_{lk}}$ and $\mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}$ are both diagonal matrices. Due to the scalable-system assumption $T = O(L)$ and the elements in $\mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}}$ being asymptotically upper and lower bounded, one can show $\mathbf{C}_{\Psi_{3,lg};\mathbf{h}_{lk}} = O(\frac{\mathbf{I}}{L})$ is upper bounded, and $\mathbf{m}_{\Psi_{3,lg};\mathbf{h}_{lk}} = O(\frac{1}{\sqrt{L}})$ according to (7.20)-(7.23). We can then show $\mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} = O(\frac{\mathbf{I}}{L})$ and $\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} = O(\frac{1}{\sqrt{L}})$ according to (7.4). $\square$

7.5 Simplification of the Messages

We define beliefs at the variable nodes as

$$\begin{aligned} b_{x_{kt}}(x_{kt}) &\propto p(x_{kt}) \prod_l \mu_{\Psi_{2,lt};x_{kt}}(x_{kt}) \\ b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk}) &\propto \mu_{\Psi_{3,lg};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) \prod_t \mu_{\Psi_{2,lt};\mathbf{h}_{lk}}(\mathbf{h}_{lk}). \end{aligned} \quad (7.27)$$

Compared to the extrinsic messages in (7.4), the beliefs in (7.27) only differ by one factor. Since Loopy BP is used for estimating x_{kt} , it has been shown in [23] that we can assume $b_{x_{kt}}(x_{kt}) \simeq \mu_{x_{kt};\Psi_{2,lt}}(x_{kt})$.

This work estimates the channel coefficients $\mathbf{h}_{lk}$ using EP. Therefore, a separate analysis from [23] is needed. We investigate the mean and covariance matrix difference between $b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})$ and $\mu_{\mathbf{h}_{lk};\Psi_{2,lt}}$ based on the Lemma 2.

Substitute (7.4) into (7.27), and we obtain the belief at $\mathbf{h}_{lk}$ as

$$b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk}) \propto \mu_{\mathbf{h}_{lk};\Psi_{2,lt}}(\mathbf{h}_{lk}) \mu_{\Psi_{2,lt};\mathbf{h}_{lk}}(\mathbf{h}_{lk}). \quad (7.28)$$

Denote $\mathbf{m}_{\hat{\mathbf{h}}_{lk}}$ and $\mathbf{C}_{\hat{\mathbf{h}}_{lk}}$ as the mean and covariance matrix of $b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})$. From Lemma 2, we have

$$\begin{aligned} \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} - \mathbf{C}_{\hat{\mathbf{h}}_{lk}} &= \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}^2 (\mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} + \mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}})^{-1} \\ \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} - \mathbf{m}_{\hat{\mathbf{h}}_{lk}} &= \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} (\mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} + \mathbf{C}_{\Psi_{2,lt};\mathbf{h}_{lk}})^{-1} \\ &\quad \cdot (\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} - \mathbf{m}_{\Psi_{2,lt};\mathbf{h}_{lk}}). \end{aligned} \quad (7.29)$$

It has been shown in the proof of Lemma 2 that the difference $\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} - \mathbf{m}_{\Psi_{2,lt};\mathbf{h}_{lk}}$ is a higher order infinitesimal relative to $\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}$. Thus, the quotients $(\mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}} - \mathbf{C}_{\hat{\mathbf{h}}_{lk}}) \mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}^{-1}$ and $(\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}} - \mathbf{m}_{\hat{\mathbf{h}}_{lk}}) / \mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}$ tend to zero as the system grows larger. Therefore, the difference in (7.29) are higher order infinitesimals relative to $\mathbf{C}_{\mathbf{h}_{lk};\Psi_{2,lt}}$ and $\mathbf{m}_{\mathbf{h}_{lk};\Psi_{2,lt}}$, respectively. Therefore, we have $b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk}) \simeq \mu_{\mathbf{h}_{lk};\Psi_{2,lt}}(\mathbf{h}_{lk})$. Based on the above discussion, we propose to replace the extrinsic (7.4) at $\Psi_{2,lt}$ by (7.30) and (7.31) to reduce complexity further,

$$\mu'_{\mathbf{h}_{lk};\Psi_{2,lt}}(\mathbf{h}_{lk}) = b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk}); \quad (7.30)$$

$$\mu'_{x_{kt};\Psi_{2,lt}}(x_{kt}) = b_{x_{kt}}(x_{kt}). \quad (7.31)$$

7.6 Decentralized Method

To obtain the belief of x_{kt} , we need to combine the message from all the AP. We consider the case where all the L AP are connected, and the AP network has a tree structure. A decentralized message-passing method can be used based on the consensus propagation framework [26]. Define the normalized message from AP l to AP l' :

$$\nu_{l \rightarrow l'}(x_{kt}) \propto \mu_{\Psi_{2,lt};x_{kt}}(x_{kt}) \prod_{\bar{l} \in N(l)/\{l'\}} \nu_{\bar{l} \rightarrow l}(x_{kt}),$$

where $N(l)$ denotes the set of connected neighbors of AP l . At convergence, the belief in (7.27) can be obtained by any AP l as

$$b'_{x_{kt}}(x_{kt}) \propto p(x_{kt}) \mu_{\Psi_{2,lt};x_{kt}}(x_{kt}) \prod_{l' \in N(l)} \nu_{l' \rightarrow l}(x_{kt}). \quad (7.32)$$

Algorithm 7.1: One Iteration of Decentralized EP

Require: $\Xi_{\mathbf{h}_{lk}}, \mathbf{y}_{p,lg}, \mathbf{y}_{lt}, p(x_{kt}), \sigma_x^2, \sigma_v^2, G_g$

- 1: Initialize $\mu_{\Psi_{3,lg};\mathbf{h}_{lk}}, \mu_{\Psi_{2,lt};x_{kt}}, \mu_{\Psi_{2,lt};\mathbf{h}_{lk}}, \nu_{l \rightarrow l'}(x_{kt})$
- 2: At all the APs, $\forall k, t$, update $b_{x_{kt}}$ according to (7.32)
- 3: **for** $l=1:L$ **do**
- 4: $\forall k$, update $\mu_{\mathbf{h}_{lk};\Psi_{3,lg}}$ based on (7.17)
- 5: $\forall k, t$, update $\mu'_{\mathbf{h}_{lk};\Psi_{2,lt}}$ based on (7.27) and (7.31)
- 6: $\forall k, t$, update $\mu'_{x_{kt};\Psi_{2,lt}}$ based on (7.30)
- 7: $\forall k$, update $\mu_{\Psi_{3,lg};\mathbf{h}_{lk}}$ based on (7.22)-(7.23)
- 8: $\forall k, t$, update $\mu_{\Psi_{2,lt};x_{kt}}$ based on (7.8)-(7.10)
- 9: $\forall k, t$, update $\mu_{\Psi_{2,lt};\mathbf{h}_{lk}}$ based on (7.14)-(7.16)
- 10: $\forall l' \in N(l), k, t$, update $\nu_{l \rightarrow l'}(x_{kt})$ based on (7.33)
- 11: **end for**

Therefore, for a decentralized algorithm, we can replace the update of belief $b_{x_{kt}}$ in (7.27) by $b'_{x_{kt}}$ in (7.32). After updating the message $\mu_{\Psi_{2,lt};x_{kt}}$, we update the shared message by

$$\nu_{l \rightarrow l'}^{new}(x_{kt}) \propto \mu_{\Psi_{2,lt};x_{kt}}^{new}(x_{kt}) \prod_{\bar{l} \in N(l)/\{l'\}} \nu_{\bar{l} \rightarrow l}^{old}(x_{kt}), \quad (7.33)$$

where we use *new* and *old* to distinguish the message of different iterations. One possible ordering method is suggested in Algorithm 7.1.

7.7 Simulation Results

Our study simulates an environment within a 400×400 square meter area, equipped with 16 APs and 8 User Terminals (UTs). Each AP features $N = 2$ antennas and is positioned at coordinates $(\frac{400}{3}i, \frac{400}{3}j)$, $i, j \in \{0, 1, 2, 3\}$. The UTs are uniformly distributed throughout the area. We denote the distance between each UT k and AP l as d_{lk} . Channel covariances for each user k at AP l are modeled using $N \times N$ diagonal matrices, represented as $\sigma_{h_{lk}}^2 \mathbf{I}$, where $10 \log_{10}(\sigma_{h_{lk}}^2) = -30 - 36.7 \log_{10}(d_{lk})$.

All the neighboring APs within $\frac{400}{3}$ meters are connected and can exchange information of the estimated data symbols. Furthermore, as illustrated in Algorithm 7.1, a synchronized message-exchanging scheme is used.

The length of the orthogonal pilot sequences is set to $P = 6$ to introduce pilot contamination.

We employ a 4QAM constellation of length $T = 10$ for signal transmission and assume a noise power of -96 dBm. The signal-to-noise ratio (SNR) is adjusted by varying the transmitted power. We base our results on 100 different realizations, which are illustrated in Figure 7.2. The normalized mean squared error (NMSE) of the channel estimates is defined as $\text{NMSE} = \frac{\text{tr}[(\widehat{\mathbf{H}} - \mathbf{H})^2]}{\text{tr}[\mathbf{H}^2]}$, where $\widehat{\mathbf{H}}$ are synthesized from the mean of $b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})$ defined in (7.27) and the operation $|\cdot|^2$ is defined as $|\mathbf{H}|^2 = \mathbf{H}^H \mathbf{H}$.

In the VL-EP scenario, we generate data symbols drawn from i.i.d. Gaussian distribution and apply the VL-EP algorithm [20] for channel estimation. In the Genie-Aided scenario, we implement the proposed algorithm as if the data symbols are known. In the MMSE Genie-Aided scenario, all the APs estimate the channel coefficients using the MMSE estimator with known channel coefficients.

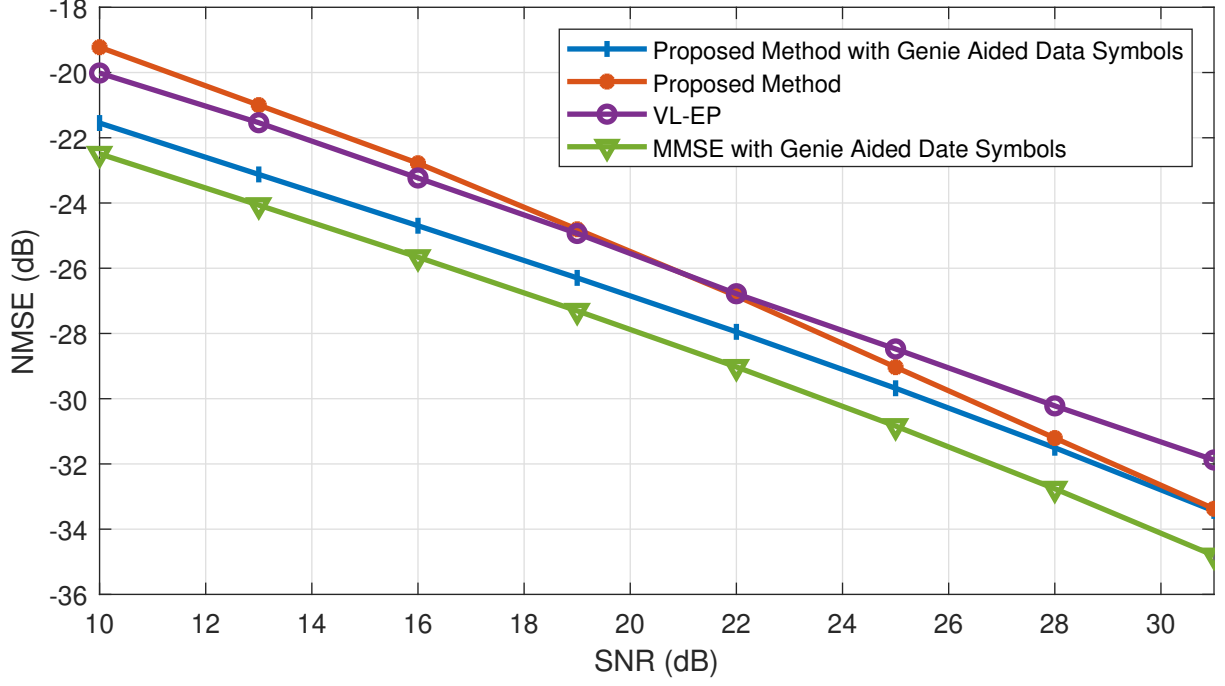


Figure 7.2: NMSE vs SNR

7.8 Conclusions

This chapter introduces a simplified, decentralized EP-based algorithm for bilinear joint estimation. To simplify the factorization scheme, we leverage orthogonal pilots and the CLT. Through asymptotic analysis, we further refine the message update scheme within the algorithm. Although originally developed for an acyclic network of APs, our simulation results confirm the algorithm's effectiveness even when the APs are interconnected in a cyclic network. *

Chapter 8

Decentralized Message-Passing for Semi-Blind Channel Estimation in Cell-Free Systems Based on Bethe Free Energy Optimization

8.1 Introduction

We examine the uplink cell-free semi-blind network containing K single-antenna user terminals (UTs) and L access points (APs). Each AP is equipped with M antennas. The received signals of the l -th AP is

$$[\mathbf{Y}_{p,l} \quad \mathbf{Y}_l] = \mathbf{H}_l [\mathbf{X}_p^\top \quad \mathbf{X}^\top] + [\mathbf{V}_{p,l} \quad \mathbf{V}_l] \in \mathbb{C}^{M \times (P+T)},$$

where the channel matrix $\mathbf{H}_l \in \mathbb{C}^{M \times K}$ comprises of independent columns. We use $\mathbf{h}_{lk}$ to denote the k -th column which follows $\mathcal{CN}(\mathbf{h}_{lk}|\mathbf{0}, \mathbf{\Xi}_{\mathbf{h}_{lk}})$. The matrix $\mathbf{X}_p$ represents the transmitted pilots. We assume that orthogonal pilots are used, i.e., difference columns of $\mathbf{X}_p$ are either the same or orthogonal. We further assume that each pilot sequence has length P and a total power of $P\sigma_x^2$. Similarly, $\mathbf{X}$ represents the data sequence. We denote the data sequence sent by the k -th UT as $\mathbf{x}_k$, which is the k -th column of $\mathbf{X}$. Moreover, we assume that each element of $\mathbf{X}$ follows an i.i.d. discrete distribution, i.e., the symbols in $\mathbf{X}$ are drawn from a constellation set $\mathcal{S}$ with power σ_x^2 . We define $\mathbf{x}_k \sim p_{\mathbf{x}_k}(\mathbf{x}_k)$. We also assume additive white Gaussian noise $\mathbf{V}_{p,l}$ and $\mathbf{V}_l$, where each entry has a power of σ_v^2 . For simplicity, we define $\mathbf{C}_v = \sigma_v^2 \mathbf{I}$.

Since the channels of different APs are independent and all the noise symbols are independent, the received signals $\mathbf{Y}_{p,l}$ and $\mathbf{Y}_l$ of different APs are conditionally independent given $\mathbf{X}$. This chapter aims to estimate $\mathbf{X}$ and $\forall l, \mathbf{H}_l$ jointly.

8.1.1 Orthogonal Pilots

When orthogonal pilots are used, we correlate the received pilot signals $\mathbf{Y}_{p,l}$ with the g -th pilot sequence $\tilde{\mathbf{x}}_{p,g}$ (not to confuse with the pilot sequence of the g -th user) to obtain the correlated version of the received pilot signals $\tilde{\mathbf{y}}_{p,lg}$:

$$\tilde{\mathbf{y}}_{p,lg} = \mathbf{Y}_{p,l} \tilde{\mathbf{x}}_{p,g}^* = P\sigma_x^2 \mathbf{H}_{lG_g} \mathbf{1}_{|G_g|} + \tilde{\mathbf{v}}_{p,lg}, \quad (8.1)$$

where we use G_g to denote the UTs groups using the g -th pilot sequence. The columns of $\mathbf{H}_{lG_g}$ are composed of the channel coefficients corresponding to the users using the g -th pilot, i.e., $\mathbf{h}_{l,k}$ is a column of $\mathbf{H}_{l,g}$ if $\mathbf{x}_{p,k} = \mathbf{x}_{p,g}$. We denote $\tilde{\mathbf{v}}_{p,l,g} = \mathbf{V}_l \tilde{\mathbf{x}}_{p,g}^*$ which is the transformed noise following a distribution $\mathcal{CN}(\mathbf{0}, \sigma_x^2 \sigma_v^2 P \mathbf{I}_M)$.

8.1.2 Factored Joint Distribution

We introduce an auxiliary variable $\mathbf{Z}_{lk} = \mathbf{h}_{lk} \mathbf{x}_k^\top$ and its vectorization $\mathbf{z}_{lk} = \text{vec}(\mathbf{Z}_{lk})$. Therefore, the likelihood of $\mathbf{Z}_{lk}$ is captured by Dirac function $p(\mathbf{Z}_{lk} | \mathbf{h}_{lk}, \mathbf{x}_k) = \delta(\mathbf{Z}_{lk} - \mathbf{h}_{lk} \mathbf{x}_k^\top)$. The joint probability density function (PDF) can be derived as

$$\begin{aligned} p(\mathbf{Y}_p, \mathbf{Y}, \mathbf{Z}_{11}, \dots, \mathbf{Z}_{LK}, \mathbf{H}_1, \dots, \mathbf{H}_L, \mathbf{X}) \\ = \prod_l p(\mathbf{Y}_l | \mathbf{Z}_{l1}, \dots, \mathbf{Z}_{lK}) \prod_l \prod_k p(\mathbf{Z}_{lk} | \mathbf{h}_{lk}, \mathbf{x}_k) \\ \prod_l \prod_g p(\tilde{\mathbf{y}}_{p,l,g}, \mathbf{H}_{lg}) \prod_k p(\mathbf{x}_k). \end{aligned} \quad (8.2)$$

For simplicity, we define

$$\begin{aligned} f_{\mathbf{z}_l}(\mathbf{z}_{l1}, \dots, \mathbf{z}_{lK}) &= p(\mathbf{Y}_l | \mathbf{Z}_{l1}, \dots, \mathbf{Z}_{lK}) \\ f_{\mathbf{h}_{lG_g}}(\mathbf{h}_{lG_g}) &= p(\tilde{\mathbf{y}}_{p,l,g}, \mathbf{H}_{lg}) \\ f_{\mathbf{x}_k}(\mathbf{x}_k) &= p(\mathbf{x}_k) \\ f_{\delta_{lk}}(\mathbf{z}_{lk}, \mathbf{h}_{lk}, \mathbf{x}_k) &= p(\mathbf{Z}_{lk} | \mathbf{h}_{lk}, \mathbf{x}_k). \end{aligned} \quad (8.3)$$

The factorization given by (8.2) admits a factor graph [5]. In this chapter, we denote $\mathbb{F} = \{f_{\mathbf{z}_l}, f_{\mathbf{h}_{lG_g}}, f_{\mathbf{x}_k}, \delta_{lk}\}$ as the set of all factor nodes and $\mathbb{V} = \{\mathbf{z}_{lk}, \mathbf{h}_{lk}, \mathbf{x}_k\}$ as the set of all variable nodes.

8.1.3 Notations

Throughout the context, we will use bold uppercase letters to denote matrices and bold lowercase letters to denote vectors. Furthermore, we use lowercase letters to denote the vectorization of uppercase letters. For example, $\mathbf{z}_{lk} = \text{vec}(\mathbf{Z}_{lk})$.

8.1.4 Prior Work

Bayesian estimation in semi-blind structures holds great potential [50], but it also presents challenges due to the high-dimensional, intractable integrals involved. Message-passing algorithms, especially expectation propagation (EP) [9], have been widely used in the field of Bayesian estimation. To enhance performance, variable-level EP (VL-EP) was proposed by combining expectation-maximization (EM) and EP [20]. To improve the convergence properties of VL-EP, hybrid EM-EP and loop-free EM-EP were introduced [55]. However, these algorithms are not equipped to handle user symbols from a finite alphabet.

To address this limitation, a distributed bilinear-EP algorithm was proposed in [2], which uses a brute-force approach to tackle the inference problem involving a finite alphabet. Inspired by [2] and [23], the authors of [56] proposed a simplified decentralized bilinear-EP algorithm.

Additionally, it has been shown [6] that various message-passing algorithms can be derived by optimizing Bethe free energy (BFE) under different assumptions. Hybrid vector message passing (HVMP) [21] was proposed based on BFE optimization, but its complexity is high, and it is unable to handle inputs from a finite alphabet.

8.1.5 Main Contributions

We propose a low-complexity algorithm for semi-blind channel and data estimation, based on the framework of BFE-constrained optimization. To effectively handle interference terms, we introduce an auxiliary variable. Furthermore, we make specific assumptions in the BFE formulation to avoid non-analytical integrals, simplifying the derivation and reducing computational complexity. Our method also integrates seamlessly with the decentralization scheme presented in [56].

In summary, the proposed algorithm blends key elements from belief propagation, expectation propagation (EP), and variational Bayes, offering a more efficient and practical solution.

8.2 Bethe Free Energy Optimization Framework

Bethe free energy is the approximated variational free energy between the true probability (8.2) and a constrained Bethe approximation trial function.

8.2.1 Bethe Approximation with Constraints

Following [6], the BFE of (8.2) is:

$$\begin{aligned}
\text{BFE} = & \sum_l D[b_{f_{z_l}}(\mathbf{z}_{l1}, \dots, \mathbf{z}_{lK}) \| p(\mathbf{Y}_l | \mathbf{z}_{l1}, \dots, \mathbf{z}_{lK})] \\
& + \sum_{l,g} D[b_{f_{\mathbf{h}_{lG_g}}}(\mathbf{h}_{lG_g}) \| p(\tilde{\mathbf{y}}_{p,lg}, \mathbf{h}_{lG_g})] + \sum_k D[b_{f_{\mathbf{x}_k}}(\mathbf{x}_k) \| p(\mathbf{x}_k)] \\
& + \sum_{l,k} D[b_{\delta_{lk}}(\mathbf{z}_{lk}, \mathbf{h}_{lk}, \mathbf{x}_k) \| \delta(\mathbf{Z}_{lk} - \mathbf{h}_{lk} \mathbf{x}_k^T)] + \sum_{l,k} H[b_{z_{lk}}(\mathbf{z}_{lk})] \\
& + \sum_{l,k} H[b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] + \sum_k L \cdot H[b_{\mathbf{x}_k}(\mathbf{x}_k)]. \tag{8.4}
\end{aligned}$$

where all the factor-level beliefs $b_{f_{z_l}}$, $b_{\delta_{lk}}$, $b_{f_{\mathbf{h}_{lG_g}}}$, $b_{f_{\mathbf{x}_k}}$, and variable-level beliefs $b_{\mathbf{h}_{lk}}$, $b_{z_{lk}}$, $b_{\mathbf{x}_k}$ are proper distributions normalized to one. Furthermore, to make all these factors consistent, the variable-level beliefs must be the marginal distribution of the factor-level beliefs. For all $l \in [1, L]$, $k \in [1, K]$, the constraints for the $\mathbf{x}_k$ are

$$\int b_{\delta_{lk}}(\mathbf{z}_{lk}, \mathbf{h}_{lk}, \mathbf{x}_k) d\mathbf{z}_{lk} d\mathbf{h}_{lk} = b_{\mathbf{x}_k}(\mathbf{x}_k) \tag{8.5}$$

$$b_{f_{\mathbf{x}_k}}(\mathbf{x}_k) = b_{\mathbf{x}_k}(\mathbf{x}_k). \tag{8.6}$$

However, satisfying the strict constraints of $\mathbf{h}_{lk}$ and $\mathbf{z}_{lk}$ will lead to an intractable problem. Therefore, we relax the strict constraints to first and second-order moment constraints

(specifically, mean and covariance constraints). W.l.o.g., we denote those sufficient statistics as $\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})$, $\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})$

$$\mathbb{E}_{b_{f_{\mathbf{z}_l}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] = \mathbb{E}_{b_{\mathbf{z}_{lk}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] \quad (8.7)$$

$$\mathbb{E}_{\delta_{lk}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] = \mathbb{E}_{b_{\mathbf{z}_{lk}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] \quad (8.8)$$

$$\mathbb{E}_{b_{f_{\mathbf{h}_{lk}G_g}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] = \mathbb{E}_{b_{\mathbf{h}_{lk}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] \quad (8.9)$$

$$\mathbb{E}_{b_{\delta_{lk}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] = \mathbb{E}_{b_{\mathbf{h}_{lk}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] \quad (8.10)$$

Moreover, to make the further derivation tractable with finite input $\mathbf{X}$, we only consider the covariance constraints of elements within every size- M block $\forall t \in [1, T]$, $[\mathbf{z}_{lk}]_{(t-1)M+1:tM}$.

8.2.2 Bethe Free Energy Optimization

The optimization criteria can be concluded by

$$\begin{aligned} & \min_b \text{BFE} \\ & \text{s.t. } (8.5) \sim (8.10). \end{aligned} \quad (8.11)$$

We observe the term $D[b_{\delta_{lk}}(\mathbf{z}_{lk}, \mathbf{h}_{lk}, \mathbf{x}_k) \|\delta(\mathbf{Z}_{lk} - \mathbf{h}_{lk}\mathbf{x}_k^\top)]$ in (8.4). Since we need to minimize the BFE, the posterior factor $b_{\delta_{lk}}$ must contain the factor $\delta(\mathbf{Z}_{lk} - \mathbf{h}_{lk}\mathbf{x}_k^\top)$ to avoid infinity BFE value. In order to have an analytical algorithm, we further assume the following form for the distribution factor $b_{\delta_{lk}}$:

$$b_{\delta_{lk}}(\mathbf{z}_{lk}, \mathbf{h}_{lk}, \mathbf{x}_k) = b_{\delta_{\mathbf{h},lk}}(\mathbf{h}_{lk})b_{\delta_{\mathbf{x},lk}}(\mathbf{x}_k)\delta(\mathbf{Z}_{lk} - \mathbf{h}_{lk}\mathbf{x}_k^\top), \quad (8.12)$$

where the belief $b_{\delta_{\mathbf{h},lk}}$ and $b_{\delta_{\mathbf{x},lk}}$ are beliefs normalized to one. By using Lagrangian methods, we can obtain the following message-passing style system of equations along with (8.12):

$$b_{f_{\mathbf{z}_l}}(\mathbf{z}_{l1}, \dots, \mathbf{z}_{lK}) = p(\mathbf{Y}_l | \mathbf{z}_{l1}, \dots, \mathbf{z}_{lK}) \prod_k \mu_{\mathbf{z}_{lk}; f_{\mathbf{z}_l}}(\mathbf{z}_{lk}) \quad (8.13)$$

$$b_{f_{\mathbf{h}_{lk}G_g}}(\mathbf{h}_{lkG_g}) = p(\tilde{\mathbf{y}}_{p,lg}, \mathbf{h}_{lkG_g}) \prod_{k \in G_g} \mu_{\mathbf{h}_{lk}; f_{\mathbf{h}_{lk}G_g}}(\mathbf{h}_{lk}) \quad (8.14)$$

$$b_{f_{\mathbf{x}_k}}(\mathbf{x}_k) = p(\mathbf{x}_k) \mu_{\mathbf{x}_k; f_{\mathbf{x}_k}}(\mathbf{x}_k) \quad (8.15)$$

$$b_{\delta_{\mathbf{h},lk}}(\mathbf{h}_{lk}) = \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}) e^{\int \delta_{\mathbf{x},lk}(\mathbf{x}_k) \ln \mu_{\mathbf{z}_{lk}; \delta_{lk}}(\text{vec}(\mathbf{h}_{lk}\mathbf{x}_k^\top)) d\mathbf{x}_k} \quad (8.16)$$

$$b_{\delta_{\mathbf{x},lk}}(\mathbf{x}_k) = \mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k) e^{\int \delta_{\mathbf{h},lk}(\mathbf{h}_{lk}) \ln \mu_{\mathbf{z}_{lk}; \delta_{lk}}(\text{vec}(\mathbf{h}_{lk}\mathbf{x}_k^\top)) d\mathbf{h}_{lk}} \quad (8.17)$$

$$b_{\mathbf{z}_{lk}}(\mathbf{z}_{lk}) = \mu_{\mathbf{z}_{lk}; f_{\mathbf{z}_l}}(\mathbf{z}_{lk}) \mu_{\mathbf{z}_{lk}; \delta_{lk}}(\mathbf{z}_{lk}) \quad (8.18)$$

$$b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk}) = \mu_{\mathbf{h}_{lk}; f_{\mathbf{h}_{lk}G_g}}(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}) \quad (8.19)$$

$$b_{\mathbf{x}_k}(\mathbf{x}_k) = [\mu_{\mathbf{x}_k; f_{\mathbf{x}_k}}(\mathbf{x}_k) \prod_l \mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k)]^{1/L}, \quad (8.20)$$

The equations (8.12)~(8.17) describes the factor level beliefs while (8.18)~(8.20) are variable level beliefs. For all $f \in \mathbb{F}$, $\theta \in \mathbb{V}$, we interpret $\mu_{\theta;f}$ as the variable to factor message. Furthermore, we can define the factor to variable messages such that the following relation holds [11]

$$\forall f \in N(\theta), \mu_{\theta;f}(\theta) = \prod_{f' \in N(\theta)/\{f\}} \mu_{f';\theta}(\theta), \quad (8.21)$$

where $N(\boldsymbol{\theta})$ denotes the neighborhood around $\boldsymbol{\theta}$ in the corresponding factor graph. Thus, (8.20) can be rewritten into the message passing form

$$b_{\mathbf{x}_k}(\mathbf{x}_k) = \mu_{f_{\mathbf{x}_k}; \mathbf{x}_k}(\mathbf{x}_k) \prod_l \mu_{\delta_{lk}; \mathbf{x}_k}(\mathbf{x}_k) \quad (8.22)$$

Since the sufficient statistics we consider here are first and second-order moments, the messages $\mu_{f_{\mathbf{h}_{lG_g}}; \mathbf{h}_{lk}}$, $\mu_{\delta_{lk}; \mathbf{h}_{lk}}$, $\mu_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}}$ and $\mu_{\delta_{lk}; \mathbf{z}_{lk}}$ are all (unnormalized) Gaussian distributions. Therefore, in the following, for all $f \in \mathbb{F}$, $\boldsymbol{\theta} \in \mathbb{V}$, we use $\mathbf{m}_{f; \boldsymbol{\theta}}$, $\mathbf{C}_{f; \boldsymbol{\theta}}$ to denote the mean and covariance of the factor-to-variable (normalized) message distributions $\mu_{f; \boldsymbol{\theta}}$. For convenience, we also denote the mean and covariance of the variable-to-factor (normalized) message $\mu_{\boldsymbol{\theta}; f}$ as $\mathbf{m}_{\boldsymbol{\theta}; f}$ and $\mathbf{C}_{\boldsymbol{\theta}; f}$. We should note here that the factor-to-variable messages fully determine those variable-to-factor messages and beliefs.

Furthermore, since the second-order sufficient statistics of $\mathbf{z}_{lk}$ only include the covariance between the elements within each block sub-vector $\forall t \in [1, T]$, $[\mathbf{z}_{lk}]_{(t-1)M+1:tM}$, the covariance matrices $\mathbf{C}_{\delta_{lk}; \mathbf{z}_{lk}} \in \mathbb{C}^{MT \times MT}$ and $\mathbf{C}_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}} \in \mathbb{C}^{MT \times MT}$ are block diagonal matrix with block size equals to M . For simplicity, for all block matrix $\mathbf{C}$, we use the notations $\{\mathbf{C}\}_{tt, M}$ to denote the t -th $M \times M$ block matrix on the diagonal of $\mathbf{C}$. Analogously, we use the notation $\{\mathbf{m}\}_{t, M}$ to denote the t -th block vector of size $M \times 1$ in $\mathbf{m}$, i.e., the subvector $[\mathbf{m}]_{(t-1)M+1:tM}$.

The computation of factor-level beliefs (8.13)~(8.15), are composed of two types of factors, the true factors given by the joint pdf model, e.g., $p(\tilde{\mathbf{y}}_{p, lg}, \mathbf{h}_{lG_g})$, and variable-to-factor messages, e.g., $\mu_{\mathbf{h}_{lk}; f_{\mathbf{h}_{lG_g}}}$. We will use the term "intrinsic" to denote the true factors and use "extrinsic" to denote the variable-to-factor messages. Those messages can be understood as the "rest" part of the approximated posteriors besides the true intrinsic. For example, if we look at (8.14), the extrinsic $\prod_{k \in G_g} \mu_{\mathbf{h}_{lk}; f_{\mathbf{h}_{lG_g}}}$ can be interpreted as an approximation of $p(\mathbf{Y}_p, \mathbf{Y}, \mathbf{h}_{lG_g})/p(\tilde{\mathbf{y}}_{p, lg}, \mathbf{h}_{lG_g})$, where $p(\mathbf{Y}_p, \mathbf{Y}, \mathbf{h}_{lG_g})$ is the marginalization result of (8.2).

Since the optimal point of the BFE can be purely represented by those factor-to-variable messages, we will focus on deriving the update of the factors-to-variable messages in the following context. Meanwhile, the update of all the variable-to-factor messages follows (8.21).

8.3 Detailed Derivations

The messages are updated iteratively (update one message a time while considering the other messages to be known) by satisfying the constraints (8.5)~(8.10), which describe the consistencies between the factor-level beliefs $\forall f \in \mathbb{F}, b_f$ and variable-level beliefs $\forall \boldsymbol{\theta} \in \mathbb{V}, b_{\boldsymbol{\theta}}$. Each pair of the (marginalized) factor-level belief and variable-level belief constrained by (8.5)~(8.10) always has one message different. We will update that different message by considering the consistency constraints. Note, in this chapter, we base our discussion on the factor-to-variable messages since the variable-to-factor message is entirely determined by the definition (8.21). We can consider the variable-to-factor messages as aliases of the corresponding factor-to-variable messages. For example, $\mu_{\mathbf{z}_{lk}; \delta_{lk}}$ is considered as the same message as $\mu_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}}$.

8.3.1 Update of $\mu_{f_{z_l}; z_{lk}}$

We first investigate the constraint between the beliefs $b_{f_{z_l}}$ and $b_{z_{lk}}$ given by (8.13) and (8.18). According to the constraint given by (8.7), we need to match the marginal mean and covariance matrix of z_{lk} . Since the pdf $p(\mathbf{y}|\mathbf{z}_{l1}, \dots, \mathbf{z}_{lK})$ is a Gaussian pdf with block-diagonal covariance matrix, the belief $b_{f_{z_l}}$ is also a Gaussian with block-diagonal covariance matrix. Because Gaussian pdf is fully determined by mean and covariance matrix, matching the moment of z_{lk} between (8.13) and (8.18) is equivalent to matching the entire distribution between the marginalized version of (8.13) $b_{f_{z_l}}(z_{lk})$ and the belief $b_{z_{lk}}(z_{lk})$ given by (8.18). Therefore, by forcing the equality $b_{f_{z_l}}(z_{lk}) = b_{z_{lk}}(z_{lk})$, the update equation for the message $\mu_{f_{z_l}; z_{lk}}$ can be obtained by Gaussian reproduction lemma [15]:

$$\mu_{f_{z_l}; z_{lk}}(z_{lk}) = \mathcal{CN}(z_{lk}|\mathbf{y}_l - \sum_{k' \neq k} \mathbf{m}_{z_{lk'}; f_{z_l}}, \mathbf{C}_v + \sum_{k' \neq k} \mathbf{C}_{z_{lk'}; f_{z_l}}).$$

8.3.2 Update of $\mu_{f_{h_{lG_g}}; h_{lk}}$

The consistency constraint between (8.14) and (8.19) is given by (8.9). A detailed derivation of the update equation can be found in [56]. The mean and covariance matrices of $\mu_{f_{h_{lG_g}}; h_{lk}}(\mathbf{h}_{lk})$ are given by

$$\mathbf{C}_{f_{h_{lG_g}}; h_{lk}} = \left(\Xi_{h_{lk}}^{-1} + \mathbf{C}_{v+h_{l\bar{k}}|\mathbf{y}}^{-1} \right)^{-1} \quad (8.23)$$

$$\mathbf{m}_{f_{h_{lG_g}}; h_{lk}} = \mathbf{C}_{f_{h_{lG_g}}; h_{lk}} \mathbf{C}_{v+h_{l\bar{k}}|\mathbf{y}}^{-1} \mathbf{m}_{v+h_{l\bar{k}}|\mathbf{y}}, \quad (8.24)$$

where $\mathbf{C}_{v+h_{l\bar{k}}|\mathbf{y}}$ can be interpreted as the covariance matrix of the interference (estimated from observations $\mathbf{y}$ and prior knowledge) plus noise, and $\mathbf{m}_{v+h_{l\bar{k}}|\mathbf{y}}$ can be interpreted as the new observation with interference terms removed, i.e.,

$$\mathbf{C}_{v+h_{l\bar{k}}|\mathbf{y}} = \frac{\sigma_v^2}{\sigma_x^2 P} \mathbf{I} + \sum_{k' \in G_g / \{k\}} \mathbf{C}_{h_{lk'}|\mathbf{Y}} \quad (8.25)$$

$$\mathbf{m}_{v+h_{l\bar{k}}|\mathbf{y}} = \frac{1}{\sigma_x^2 P} \mathbf{y}_{p,lg} - \sum_{k' \in G_g / \{k\}} \mathbf{m}_{h_{lk'}|\mathbf{Y}_d}, \quad (8.26)$$

where

$$\begin{aligned} \mathbf{C}_{h_{lk}|\mathbf{Y}_d} &= (\Xi_{h_{lk}}^{-1} + \mathbf{C}_{h_{lk}; f_{h_{lG_g}}}^{-1})^{-1} \\ \mathbf{m}_{h_{lk}|\mathbf{Y}_d} &= \mathbf{C}_{h_{lk}|\mathbf{Y}_d} \mathbf{C}_{h_{lk}; f_{h_{lG_g}}}^{-1} \mathbf{m}_{h_{lk}; f_{h_{lG_g}}}. \end{aligned} \quad (8.27)$$

8.3.3 Update of $\mu_{f_{x_k}; x_k}$

The consistency constraint between (8.15) and (8.22) is given by (8.6). Thus, we can immediately get

$$\mu_{f_{x_k}; x_k}(\mathbf{x}_k) = p(\mathbf{x}_k). \quad (8.28)$$

8.3.4 Update of $\mu_{\delta_{lk}; x_k}$

The update of the message $\mu_{\delta_{lk}; x_k}$ is obtained by satisfying the consistency constraint (8.5) between the beliefs (8.17) and (8.22). Following the definition of $b_{\delta_{lk}}$ in (8.12), we can

immediately obtain $\mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k) = b_{\delta_{\mathbf{x},lk}}(\mathbf{x}_k)/\mu_{\mathbf{x}_k;\delta_{lk}}(\mathbf{x}_k)$. It can be seen from (8.16) that the belief $b_{\delta_{\mathbf{h},lk}}$ is Gaussian (more details in section 8.3.5. In fact, we can see $b_{\delta_{\mathbf{h},lk}} = b_{\mathbf{h}_{lk}}$). Thus, the message $\mu_{\delta_{lk};\mathbf{x}_k}$ can be derived as

$$\mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k) \propto \prod_t \mathcal{CN}(x_{kt} | \hat{m}_{\delta_{lk};\mathbf{x}_{kt}}, \hat{\tau}_{\delta_{lk};\mathbf{x}_{kt}}), \quad (8.29)$$

with

$$\hat{\tau}_{\delta_{lk};\mathbf{x}_{kt}} = \text{tr} \left[\{ \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}} \}_{tt,M}^{-1} \mathbf{R}_{b_{\delta_{\mathbf{h},lk}}} \right]^{-1} \quad (8.30)$$

$$\hat{m}_{\delta_{lk};\mathbf{x}_{kt}} = \hat{\tau}_{\delta_{lk};\mathbf{x}_{kt}} \mathbf{m}_{b_{\delta_{\mathbf{h},lk}}}^H \{ \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}} \}_{tt,M}^{-1} \{ \mathbf{m}_{\mathbf{z}_{lk};\delta_{lk}} \}_{t,M}, \quad (8.31)$$

where $\mathbf{m}_{b_{\delta_{\mathbf{h},lk}}}$ and $\mathbf{R}_{b_{\delta_{\mathbf{h},lk}}} = \mathbf{C}_{b_{\delta_{\mathbf{h},lk}}} + \mathbf{m}_{b_{\delta_{\mathbf{h},lk}}} \mathbf{m}_{b_{\delta_{\mathbf{h},lk}}}^H$ denote the mean and correlation matrix of the Gaussian pdf $b_{\delta_{\mathbf{h},lk}}$ calculated in section 8.3.5. Note here that the (normalized) message $\mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k)$ is a categorical distribution, and thus, the variables $\hat{m}_{\delta_{lk};\mathbf{x}_{kt}}, \hat{\tau}_{\delta_{lk};\mathbf{x}_{kt}}$ are just parameters for computing the message, they do not correspond to the mean and variance of the elements in $\mathbf{x}_k \sim \mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k)$. From this point, we can also update the belief by

$$b_{\delta_{\mathbf{x},lk}}(\mathbf{x}_k) = \mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k) \mu_{\mathbf{x}_k;\delta_{lk}}(\mathbf{x}_k). \quad (8.32)$$

8.3.5 Update of $\mu_{\delta_{lk};\mathbf{h}_{lk}}$

The beliefs given by (8.16) and (8.19) should satisfy the consistency constraint (8.10). The exponential factor in (8.16) can be verified as an (unnormalized) Gaussian. Therefore, the belief $b_{\delta_{\mathbf{h},lk}}$ is Gaussian, and the outbound message is computed by $\mu_{\delta_{lk};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) = b_{\delta_{\mathbf{h},lk}}(\mathbf{h}_{lk})/\mu_{\mathbf{h}_{lk};\delta_{lk}}(\mathbf{h}_{lk})$. The mean and covariance matrix of the Gaussian message $\mu_{\delta_{lk};\mathbf{h}_{lk}}$ are

$$\begin{aligned} \mathbf{C}_{\delta_{lk};\mathbf{h}_{lk}} &= \left(\sum_t [\mathbf{r}_{b_{\delta_{\mathbf{x},lk}}}]_t \{ \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}} \}_{tt,M}^{-1} \right)^{-1} \\ \mathbf{m}_{\delta_{lk};\mathbf{h}_{lk}} &= \mathbf{C}_{\delta_{lk};\mathbf{h}_{lk}} \left(\sum_t [\mathbf{m}_{b_{\delta_{\mathbf{x},lk}}}]_t^* \{ \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}} \}_{tt,M} \{ \mathbf{m}_{\mathbf{z}_{lk};\delta_{lk}} \}_{t,M} \right) \end{aligned}$$

where $\mathbf{r}_{b_{\delta_{\mathbf{x},lk}}} = \mathbb{E}_{b_{\delta_{\mathbf{x},lk}}} [\mathbf{x}_k \cdot \mathbf{x}_k^*]$, $\mathbf{m}_{b_{\delta_{\mathbf{x},lk}}} = \mathbb{E}_{b_{\delta_{\mathbf{x},lk}}} [\mathbf{x}_k]$ and "." denotes element-wise product. Thus, we update the belief by

$$\begin{aligned} b_{\delta_{\mathbf{h},lk}}(\mathbf{h}_{lk}) &= \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{b_{\delta_{\mathbf{h},lk}}}, \mathbf{C}_{b_{\delta_{\mathbf{h},lk}}}) \\ &= \mu_{\delta_{lk};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk};\delta_{lk}}(\mathbf{h}_{lk}). \end{aligned} \quad (8.33)$$

8.3.6 Update of $\mu_{\delta_{lk};\mathbf{z}_{lk}}$

We examine the moments consistency between (8.12) and (8.18) based on (8.8). Note here the message $\mu_{\mathbf{z}_{lk};\delta_{lk}}$ is implicitly included in $b_{\delta_{lk}}$ due to the definition (8.16) and (8.16). Therefore, we will update $\mu_{\delta_{lk};\mathbf{z}_{lk}}$ to make (8.12) and (8.18) consistent. The update equation of $\mu_{\delta_{lk};\mathbf{z}_{lk}}$ can be derived to be

$$\mu_{\delta_{lk};\mathbf{z}_{lk}}(\mathbf{z}_{lk}) = \frac{\text{proj}[b_{\delta_{lk}}(\mathbf{z}_{lk})]}{\mu_{\mathbf{z}_{lk};\delta_{lk}}(\mathbf{z}_{lk})}, \quad (8.34)$$

where the operation $q(\mathbf{z}_{lk}) = \text{proj}[p(\mathbf{z}_{lk})]$ projects the distribution p to Gaussian family q with block covariance matrices, such that the sufficient statistics $\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})$ of p and q are the same. Thus, the message $\mu_{\delta_{lk};\mathbf{z}_{lk}}$ is updated by

$$\mu_{\delta_{lk};\mathbf{z}_{lk}}(\mathbf{z}_{lk}) = \frac{\mathcal{CN}(\mathbf{z}_{lk} | \mathbf{m}_{b_{\delta_{\mathbf{z},lk}}}, \mathbf{C}_{b_{\delta_{\mathbf{z},lk}}})}{\mathcal{CN}(\mathbf{z}_{lk} | \mathbf{m}_{\mathbf{z}_{lk};\delta_{lk}}, \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}})}, \quad (8.35)$$

with

$$\begin{aligned} \mathbf{m}_{b_{\delta_{\mathbf{z},lk}}} &= \mathbf{m}_{b_{\delta_{\mathbf{x},lk}}} \otimes \mathbf{m}_{b_{\delta_{\mathbf{h},lk}}} \\ \mathbf{C}_{b_{\delta_{\mathbf{z},lk}}} &= \text{diag}(\mathbf{r}_{b_{\delta_{\mathbf{x},lk}}}) \otimes \mathbf{C}_{b_{\delta_{\mathbf{h},lk}}} + \mathbf{C}_{b_{\delta_{\mathbf{x},lk}}} \otimes \mathbf{m}_{b_{\delta_{\mathbf{h},lk}}} \mathbf{m}_{b_{\delta_{\mathbf{h},lk}}}^H, \end{aligned}$$

where $\mathbf{C}_{b_{\delta_{\mathbf{x},lk}}} = \mathbb{E}_{b_{\delta_{\mathbf{x},lk}}}[\mathbf{x}_k \mathbf{x}_k^H]$. According to (8.32), the belief $b_{\delta_{\mathbf{x},lk}}$ is entirely determined by the messages $\mu_{f_{\mathbf{x}_k};\mathbf{x}_k}$ and $\forall l, \mu_{\delta_{lk};\mathbf{x}_k}$, which are all independent according to (8.28) and (8.29). Therefore, the covariance matrix $\mathbf{C}_{b_{\delta_{\mathbf{x},lk}}}$ is a diagonal matrix.

Note that the belief distribution $b_{\delta_{lk}}(\mathbf{z}_{lk})$ is a Gaussian mixture model. Thus, the resulting covariance matrix $\mathbf{C}_{\delta_{lk};\mathbf{z}_{lk}}$ may not be positive semi-definite. We propose the following correction: whenever the eigenvalue of $\mathbf{C}_{\delta_{lk};\mathbf{z}_{lk}}$ is negative, we reset it to a large value to indicate that the estimation during the current iteration step is not correct. Since we are using the iterative algorithm to find the fixed point of the BFE, resetting the value will not change the final result.

8.4 Decentralization Method

Until this point, we have developed a distributed BFE-based message-passing algorithm since a CPU is needed to compute the messages $\mu_{\mathbf{x}_k;\delta_{lk}}$. These messages are only used to compute the beliefs $b_{\delta_{\mathbf{x},lk}}$ which are then used to update the messages $\mu_{\delta_{lk};\mathbf{h}_{lk}}$ and $\mu_{\delta_{lk};\mathbf{z}_{lk}}$. Based on this observation, we use the backhaul message-passing scheme (physical message passed from AP to AP) proposed in [56] to decentralize the computing of $b_{\delta_{\mathbf{x},lk}}$. We define the update rule of the backhaul message from AP l to AP l' to be

$$\nu_{l \rightarrow l'}(\mathbf{x}_k) = \mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k) \prod_{l'' \in N(l) \setminus \{l'\}} \nu_{l'' \rightarrow l}(\mathbf{x}_k), \quad (8.36)$$

where we exploit the notations and use $N(l)$ to denote the neighborhood of AP l in the AP network. At each AP, the approximated version of belief $b_{\delta_{\mathbf{x},lk}}(\mathbf{x}_k)$ is recovered by

$$\hat{b}_{\delta_{\mathbf{x},lk}}(\mathbf{x}_k) = p(\mathbf{x}_k) \mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k) \prod_{l' \in N(l)} \nu_{l' \rightarrow l}(\mathbf{x}_k). \quad (8.37)$$

This approximated belief is exact when the two conditions hold: 1), the backhaul messages converge to the steady point; 2), the AP network is acyclic. Nevertheless, we will use (8.37) to replace the exact update in (8.32). A suggested update order is concluded in Algorithm 8.1. As we can see, at each AP, our algorithm has a complexity of $O[(M^3 + |\mathcal{S}|)KT]$, where $|\mathcal{S}|$ denotes the size of the $\mathcal{S}$.

Algorithm 8.1: Proposed Method in one iteration

Require: $\forall l, g, k, \mathbf{y}_l, p(\mathbf{x}_k), p(\mathbf{h}_{lk}), p(\mathbf{y}_l | \mathbf{z}_{l1}, \dots, \mathbf{z}_{lK})$

- 1: Initialize: All the factor-to-variable messages and $b_{\delta_{\mathbf{h},lk}}$ s.t.: 1), If Gaussian, then zero mean and unit covariance matrices. 2), If categorical distribution, then uniform.
 - 2: $[\forall l, \text{ At AP } l, \text{ execute the following loop}]$
 - 3: **repeat** $[\forall l' \in N(l)k, g]$
 - 4: $\mu_{\mathbf{z}_{lk};f_{\mathbf{z}_l}}(\mathbf{z}_{lk}) = \mu_{\delta_{lk};\mathbf{z}_{lk}}(\mathbf{z}_{lk})$
 - 5: Update $\mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}}$ based on Section 8.3.1
 - 6: $\mu_{\mathbf{h}_{lk};f_{\mathbf{h}_{lG_g}}}(\mathbf{h}_{lk}) = \mu_{\delta_{lk};\mathbf{h}_{lk}}(\mathbf{h}_{lk})$
 - 7: Update $\mu_{f_{\mathbf{h}_{lG_g}};\mathbf{h}_{lk}}$ based on Section 8.3.2
 - 8: $\mu_{f_{\mathbf{x}_k};\mathbf{x}_k}(\mathbf{x}) = p(\mathbf{x}_k)$ due to Section 8.3.3
 - 9: $\mu_{\mathbf{z}_{lk};\delta_{lk}}(\mathbf{z}_{lk}) = \mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}}(\mathbf{z}_{lk})$
 - 10: Update $\mu_{\delta_{lk};\mathbf{x}_k}$ based on Section 8.3.4
 - 11: $\nu_{l \rightarrow l'}(\mathbf{x}_k) = \mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k) \prod_{l'' \in N(l)/\{l'\}} \nu_{l'' \rightarrow l}(\mathbf{x}_k)$
 - 12: $b_{\delta_{\mathbf{x},lk}}(\mathbf{x}_k) = p(\mathbf{x}_k) \mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k) \prod_{l' \in N(l)} \nu_{l' \rightarrow l}(\mathbf{x}_k)$
 - 13: $\mu_{\mathbf{h}_{lk};\delta_{lk}}(\mathbf{h}_{lk}) = \mu_{f_{\mathbf{h}_{lG_g}};\mathbf{h}_{lk}}(\mathbf{h}_{lk})$
 - 14: Update $\mu_{\delta_{lk};\mathbf{h}_{lk}}$ based on Section 8.3.5
 - 15: $b_{\delta_{\mathbf{h},lk}}(\mathbf{h}_{lk}) = \mu_{\delta_{lk};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk};\delta_{lk}}(\mathbf{h}_{lk})$ based on (8.33)
 - 16: Update $\mu_{\delta_{lk};\mathbf{z}_{lk}}$ based on Section 8.3.6
 - 17: **until** Convergence
-

8.5 Simulation Results

In this section, we will verify the algorithm using numerical simulations. We consider a $400m \times 400m$ area with $M = 16$ APs and $K = 8$ UTs. The APs are located at the coordinates $(\frac{400}{3}i, \frac{400}{3}j)$, where $i, j \in \{0, \dots, 3\}$. The UTs are uniformly randomly distributed over this area. The fading model we use is [2],

$$\sigma_{l,k}^2[\text{dB}] = -30.5 - 36.7 \log_{10}(d_{lk}), \quad (8.38)$$

where d_{lk} is the distance between AP l and UT k . All the neighboring APs within $\frac{400}{3}$ meters are connected and can exchange information of the estimated data symbols. Furthermore, as illustrated in Algorithm 8.1, a synchronized message-exchanging scheme is used. To induce pilot contamination, the default pilot sequence length is $P = 6$. The transmitted data sequence spans a length of $L = 10$. We use 4-QAM constellation to generate the i.i.d. input symbols $\mathbf{X}^T$.

We maintain consistent positions for all APs and UTs for different realizations and conduct simulations across 50 unique scenarios with varying $\mathbf{H}$, $\mathbf{V}$, and user data. The metric for evaluating performance is channel normalized mean squared error (NMSE). It is calculated as $\frac{\sum \text{tr}(\tilde{\mathbf{H}}\tilde{\mathbf{H}}^H)}{\sum \text{tr}(\mathbf{H}\mathbf{H}^H)}$, where $\tilde{\mathbf{H}}$ represents the estimation error. The simulation results are concluded in Fig. 8.1. For comparison, we also plot the results of the VL-EP algorithm [20], which assumes Gaussian inputs, and the EP-based decentralized algorithm [56], which assumes discrete inputs. In the Genie-Aided scenario, we assume the data to be known. The performance curve of our proposed method coincides with the EP-based decentralized method. However, EP-based decentralized has a higher complexity, i.e., $O[(|\mathcal{A}|M^3 + |\mathcal{S}|)KT]$ at each AP, where $\mathcal{A} = \{x^2 | x \in \mathcal{S}\}$.

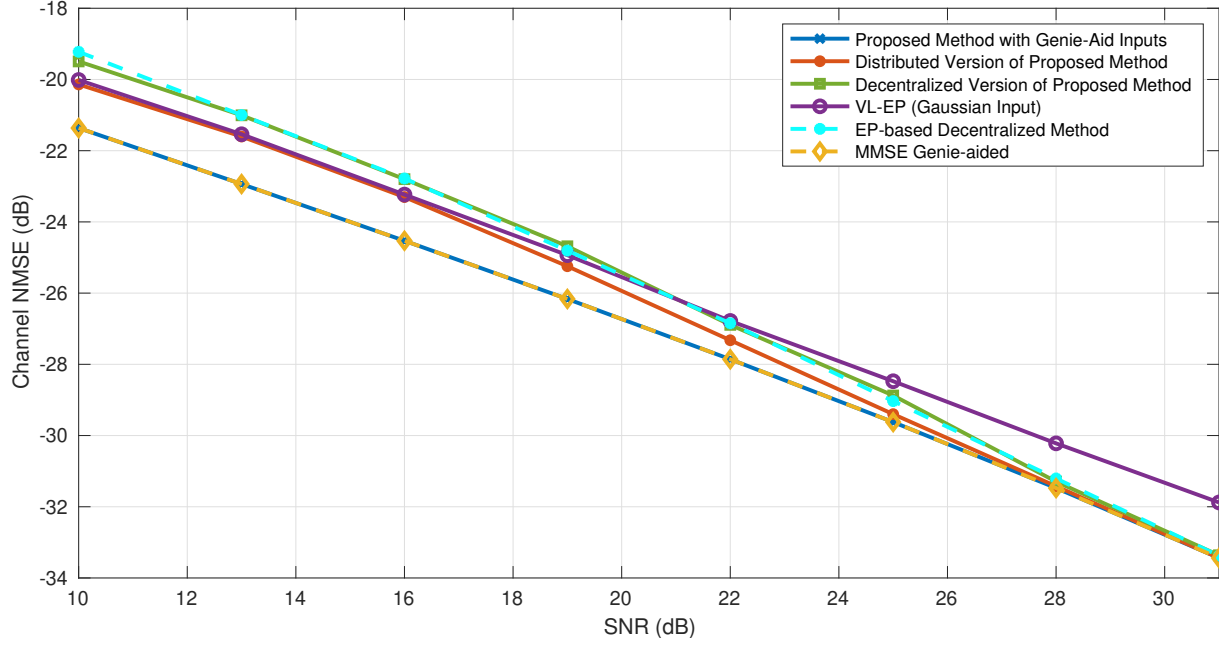


Figure 8.1: Channel NMSE (dB) versus SNR (dB)

8.6 Conclusions

In this chapter, we begin with constrained BFE optimization. To avoid non-analytical integrals, such as Gaussian mixtures, we introduce a key assumption (8.12) for the BFE. By iteratively enforcing consistency constraints, we derive the proposed low-complexity message-passing algorithm for semi-blind estimations. Since the message updates depend only on the belief $b_{\delta_{x,lk}}$, our algorithm integrates seamlessly into the decentralized scheme

Chapter 9

Hierarchical Expectation Propagation for Semi-Blind Channel Estimation in Cell-Free Networks

9.1 Introduction

Unlike conventional mobile networks, Cell-Free (CF) networks no longer have the cellular topology. As a result, all the user terminals (UTs) will be served by all the access points (APs) in an area. A huge gain throughput gain can be obtained by transforming into the new CF network topology [57]. One of the challenges introduced by CF network is pilot contamination, which occurs when the number of UTs exceeds the length of pilot sequence. As a result, multiple UTs may use the same pilot sequence.

Semi-Blind approaches [50] have been explored to mitigate the effects of pilot contamination. In Semi-Blind approaches, the channel and data are jointly estimated with the presence of pilot contamination. The authors have also shown a performance gain by transforming from deterministic estimation to Bayesian estimation [50, 51]. However, Bayesian inference is intractable due to the high dimensional integrals.

A variety of Message Passing based algorithms has been proposed to simplify the high dimensional Bayesian inference problem into local low dimensional inference problems [34, 37, 58–60].

9.1.1 Continuous Input

Variable Level Expectation Propagation (VL-EP) [20] is a variant of Expectation Propagation (EP) that assumes Gaussian input from UTs and combines EP with Expectation Maximization (EM) for joint bilinear estimation. In VL-EP, posterior distributions (referred to as beliefs in EP terminology) are used as extrinsic messages to reduce complexity. In [55], the authors investigated the differences between VL-EP and EP algorithms and proposed a block coordinate descent method based on the Majorization-Minimization (MM) algorithm. However, in more realistic scenarios, UT inputs typically follow a discrete distribution, such as Quadrature Amplitude Modulation (QAM). Unfortunately, the aforementioned methods do not perform well for discrete input cases.

9.1.2 Discrete Input

In [2], the authors proposed an EP-based distributed algorithm for bilinear estimation with discrete UT inputs. However, the complexity of this approach scales linearly with the size of the constellation set. In [61], the authors exploited the relationship between Bethe Free Energy (BFE) [58] and message-passing algorithms. By incorporating mean-field constraints into the BFE, the resulting low-complexity algorithm, proposed in [61], can be interpreted as a combination of Variational Bayes (VB) and EP.

9.1.3 Main Contributions

In this chapter, we propose an algorithm to mitigate the high complexity associated with large constellation sets. As observed in [56] and [2], differences in the amplitudes of constellation symbols further contribute to this complexity. Additionally, in decentralized or distributed algorithms, access points (APs) need to share information about data symbols through backhaul connections. This information is represented as a list of probabilities corresponding to each element in the constellation set, and its size grows linearly with the size of the constellation set. While the approach in [61] ensures that backhaul information exchanged between APs is independent of the constellation size, it still requires traversing the entire constellation set to determine the transmitted symbol.

To address these challenges, we exploit the hierarchical structure of the constellation set. Specifically, we assume that the original constellation set can be decomposed into multiple sub-constellation sets, each with identical amplitudes. A commonly used example is 4^M -QAM, which can be decomposed into M 4-QAM sub-constellation sets. By leveraging this decomposition, backhaul messages become Gaussian, requiring only the first and second-order moments to be shared between APs. Furthermore, this approach reduces the complexity of symbol determination from 4^M to $4M$ possible values per data symbol.

9.2 System Model

We consider a semi-blind signal model containing L APs and K UTs. At the l -th AP, it is represented by

$$[\mathbf{Y}_{p,l} \quad \mathbf{Y}_l] = \mathbf{H}_l [\mathbf{X}_p \quad \mathbf{X}] + [\mathbf{V}_{p,l} \quad \mathbf{V}_l]. \quad (9.1)$$

The received signals are composed of pilot part $\mathbf{Y}_{p,l} \in \mathbb{C}^{M \times P}$ and data part $\mathbf{Y}_l \in \mathbb{C}^{N \times T}$. The channels between different users are considered independent Gaussian i.e. $\text{vec}(\mathbf{H}_l) \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Xi}_l)$ where $\mathbf{\Xi}_l \in \mathbb{C}^{NK \times NK}$ is a block diagonal matrix of K blocks $\mathbf{\Xi}_{h_{lk}} \in \mathbb{C}^{N \times N}$. The transmitted symbols can be decomposed as pilot symbols $\mathbf{X}_p \in \mathcal{S}^{K \times P}$ and data symbols $\mathbf{X} \in \mathcal{S}^{K \times T}$, where $\mathcal{S}$ is the constellation set. The signal power is denoted as σ_x^2 . The noise is considered as i.i.d. Gaussian distribution, and thus, $\text{vec}([\mathbf{V}_{p,l} \quad \mathbf{V}_l]) \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_v)$ with $\mathbf{C}_v = \sigma_v^2 \mathbf{I}$.

9.2.1 Orthogonal Pilot sequences

If orthogonal pilot sequences are used, we can first preprocess the pilot observation by right multiplying it with $\tilde{\mathbf{x}}_{p,g}^*$ which is the conjugated g -th pilot sequence. This results in an equivalent observation $\tilde{\mathbf{y}}_{p,lg}$

$$\tilde{\mathbf{y}}_{p,lg} = \frac{1}{P\sigma_x^2} \mathbf{Y}_{p,l} \tilde{\mathbf{x}}_{p,g}^* = \sum_{k \in G_g} \mathbf{h}_{lk} + \tilde{\mathbf{v}}_{p,lg} \quad (9.2)$$

where $\tilde{\mathbf{v}}_{p,lg} = \frac{1}{P\sigma_x^2} \mathbf{V}_{p,l} \tilde{\mathbf{x}}_{p,g}^* \sim \mathcal{N}(\mathbf{v}_{p,lg} | \mathbf{0}, \mathbf{C}_{\tilde{\mathbf{v}}})$, $\mathbf{C}_{\tilde{\mathbf{v}}} = \frac{\sigma_v^2 \mathbf{I}}{P\sigma_x^2}$, and G_g denotes the set of users using the g -th pilot sequence. We observe that every $\mathbf{h}_{lk}$ occurs only in one group G_g , and the cross-correlation $\mathbb{E}[\tilde{\mathbf{v}}_{p,lg} \tilde{\mathbf{v}}_{p,lg'}^H]$ is an all-zero matrix for all $g \neq g'$. Therefore, the observations $\tilde{\mathbf{y}}_{p,lg}$ and $\tilde{\mathbf{y}}_{p,lg'}$ are independent.

9.2.2 Hierarchical Data Symbol

In this chapter, we assume that the data symbols can be decomposed into

$$x_{kt} = \sum_{m=1}^M x_{m,kt}, \quad (9.3)$$

where $|x_{m,kt}|^2 = \sigma_{xm}^2$. A common example of this constellation scheme is 4^M -QAM. If we assume the transmission power of 4^M -QAM to be σ_x^2 , then we have

$$\sigma_{xm}^2 = \frac{3\sigma_x^2 \frac{1}{4}^{m-1}}{4 \left(1 - \frac{1}{4}^M\right)}. \quad (9.4)$$

We define an auxiliary variable

$$\mathbf{z}_{m,lt} = \sum_k \mathbf{h}_{lk} x_{m,kt}. \quad (9.5)$$

The joint pdf can be factored as

$$\begin{aligned} & p(\{\mathbf{y}_{p,lg}\}, \{\mathbf{Y}_l\}, \{\mathbf{z}_{m,lt}\}, \{\mathbf{H}_l\}, \mathbf{X}, \{\mathbf{V}_l\}) \\ &= \prod_l \prod_t f_{\mathbf{y}_{lt}}(\mathbf{z}_{\{m\},lt}) \prod_m \prod_l \prod_t \delta_{m,lt}(\mathbf{z}_{m,lt}, \mathbf{h}_{l\{k\}}, \mathbf{x}_{m,\{k\}t}) \\ & \cdot \prod_l \prod_g f_{\mathbf{H}_{lG_g}}(\mathbf{H}_{lG_g}) \prod_m \prod_k \prod_t f_{x_{m,kt}}(x_{m,kt}) \end{aligned} \quad (9.6)$$

where

$$\begin{aligned} f_{\mathbf{y}_{lt}}(\mathbf{z}_{\{m\},lt}) &= p\left(\mathbf{y}_{lt} \mid \sum_m \mathbf{z}_{m,lt}\right); \\ \delta_{m,lt}(\mathbf{z}_{m,lt}, \mathbf{h}_{l\{k\}}, \mathbf{x}_{m,\{k\}t}) &= p\left(\mathbf{z}_{m,lt} \mid \sum_k \mathbf{h}_{lk} x_{m,kt}\right); \\ f_{\mathbf{H}_{lG_g}}(\mathbf{H}_{lG_g}) &= p(\mathbf{y}_{p,lg}, \mathbf{H}_{lG_g}); \\ f_{x_{m,kt}}(x_{m,kt}) &= p(x_{m,kt}) \end{aligned} \quad (9.7)$$

where the curly bracket in the subscript $\theta_{\{n\}}$ denotes all $\theta_1, \dots, \theta_N$, and $\mathbf{H}_{lG_g}$ denotes a matrix collecting all $k \in G_g$, $\mathbf{h}_{lk}$ as its column vectors. We will base our EP (BP) algorithm based on this factorization scheme.

9.2.3 Notations

The variable-to-factor messages are fully determined by the factor-to-variable messages. Therefore, in the following discussion, we will focus on deriving the factor-to-variable messages. In the EP procedure, we denote the mean and covariance of the variable-to-factor message $\mu_{\Psi;\theta_i}$ as $\mathbf{m}_{\Psi;\theta_i}$ and $\mathbf{C}_{\Psi;\theta_i}$, and those of the factor-to-variable message $\mu_{\theta_i;\Psi}$ as $\mathbf{m}_{\theta_i;\Psi}$ and $\mathbf{C}_{\theta_i;\Psi}$.

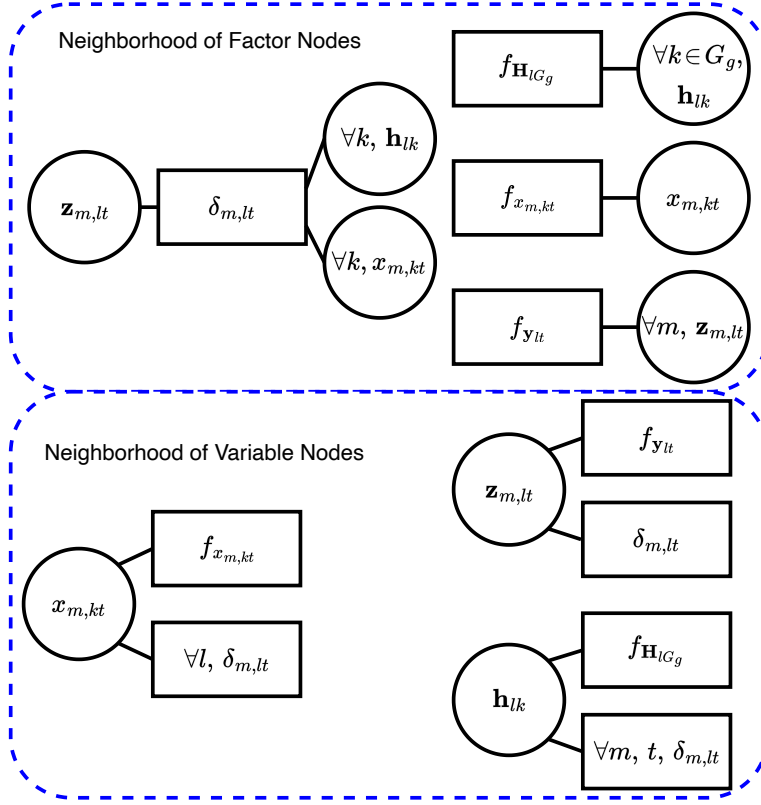


Figure 9.1: Factor Graph with Neighborhoods of both Factor (above) and Variable (below) Nodes. For simplicity, $\forall$ symbol is used to indicate that there are multiple nodes.

9.3 Message Passing Derivations

We first look at the factor node $\delta_{m,lt}$. The (partially) marginalized belief at $\delta_{m,lt}$ is

$$\begin{aligned}
& b_{\delta_{m,lt}}(\mathbf{z}_{m,lt}, \mathbf{h}_{lk}, x_{m,kt}) \\
&= \int \delta(\mathbf{z}_{m,lt} - \sum_k \mathbf{h}_{lk} x_{m,kt}) \mu_{\mathbf{z}_{m,lt}; \delta_{m,lt}}(\mathbf{z}_{m,lt}) \\
&\cdot \prod_k \mu_{\mathbf{h}_{lk}; \delta_{m,lt}}(\mathbf{h}_{lk}) \prod_k \mu_{x_{m,kt}; \delta_{m,lt}}(x_{m,kt}) d\mathbf{h}_{\bar{l}\bar{k}} dx_{m,\bar{l}\bar{k}t} \\
&= \mu_{\mathbf{z}_{m,lt}; \delta_{m,lt}}(\mathbf{z}_{m,lt}) \mu_{\mathbf{h}_{lk}; \delta_{m,lt}}(\mathbf{h}_{lk}) \mu_{x_{m,kt}; \delta_{m,lt}}(x_{m,kt}) \\
&\cdot \int d\mathbf{w}_{m,lkt} \delta(\mathbf{z}_{m,lt} - \mathbf{w}_{m,lkt} - \mathbf{h}_{lk} x_{m,kt}) \\
&\cdot \int d\mathbf{h}_{\bar{l}\bar{k}} dx_{m,\bar{l}\bar{k}t} \delta(\mathbf{w}_{m,lkt} - \sum_{k' \neq k} \mathbf{h}_{lk'} x_{m,k't}) \\
&\cdot \prod_{k' \neq k} \mu_{\mathbf{h}_{lk'}; \delta_{m,lt}}(\mathbf{h}_{lk'}) \prod_{k' \neq k} \mu_{x_{m,k't}; \delta_{m,lt}}(x_{m,k't}),
\end{aligned} \tag{9.8}$$

where we define $\mathbf{w}_{m,lkt} = \sum_{k' \neq k} \mathbf{h}_{lk'} x_{m,k't}$ as the interference in (9.8). If we look at the second integral after the second equal sign in (9.8), this interference $\mathbf{w}_{m,lkt}$ can be

approximated as Gaussian by Central Limit Theory, with mean and covariance matrix

$$\begin{aligned}\mathbf{m}_{\mathbf{w}_{m,lkt}} &= \sum_{k' \neq k} \mathbf{m}_{\mathbf{h}_{lk'}; \delta_{m,lt}} m_{x_{m,k't}; \delta_{m,lt}} \\ \mathbf{C}_{\mathbf{w}_{m,lkt}} &= \sum_{k' \neq k} \mathbf{C}_{\mathbf{h}_{lk'}; \delta_{m,lt}} \sigma_{xm}^2 \\ &\quad + \mathbf{m}_{\mathbf{h}_{lk'}; \delta_{m,lt}} \mathbf{m}_{\mathbf{h}_{lk'}; \delta_{m,lt}}^H \tau_{x_{m,k't}; \delta_{m,lt}}.\end{aligned}\tag{9.9}$$

With the above CLT approximation, the term $\mathbf{z}_{m,lt} - \mathbf{w}_{m,lkt}$ in (9.8) can be verified to be a Gaussian with mean and covariance matrix:

$$\begin{aligned}\mathbf{m}_{\mathbf{u}_{m,lkt}} &= \mathbf{m}_{\mathbf{z}_{m,lt}; \delta_{m,lt}} - \mathbf{m}_{\mathbf{w}_{m,lkt}} \\ \mathbf{C}_{\mathbf{u}_{m,lkt}} &= \mathbf{C}_{\mathbf{z}_{m,lt}; \delta_{m,lt}} + \mathbf{C}_{\mathbf{w}_{m,lkt}}\end{aligned}\tag{9.10}$$

9.3.1 Message from $\delta_{m,lt}$ to $x_{m,kt}$

From the derivation in [56] and the above CLT approximations, the feedback message $\mu_{\delta_{m,lt}; x_{m,kt}}$ can be identified as Gaussian with mean and variance

$$\begin{aligned}\tau_{\delta_{m,lt}; x_{m,kt}} &= \left[\mathbf{m}_{\mathbf{h}_{lk}; \delta_{m,lt}}^H (\mathbf{C}_{\mathbf{u}_{m,lkt}} + \sigma_{xm}^2 \mathbf{C}_{\mathbf{h}_{lk}; \delta_{m,lt}})^{-1} \cdot \mathbf{m}_{\mathbf{h}_{lk}; \delta_{m,lt}} \right]^{-1} \\ m_{\delta_{m,lt}; x_{m,kt}} &= \tau_{\delta_{m,lt}; x_{m,kt}} \mathbf{m}_{\mathbf{h}_{lk}; \delta_{m,lt}}^H \cdot (\mathbf{C}_{\mathbf{u}_{m,lkt}} + \sigma_{xm}^2 \mathbf{C}_{\mathbf{h}_{lk}; \delta_{m,lt}})^{-1} \mathbf{m}_{\mathbf{u}_{m,lkt}}.\end{aligned}\tag{9.11}$$

The belief of $x_{m,kt}$ at $\delta_{m,lt}$ factor node is

$$b_{\delta_{m,lt}; x_{m,kt}}(x_{m,kt}) = \mu_{x_{m,kt}; \delta_{m,lt}}(x_{m,kt}) \mu_{\delta_{m,lt}; x_{m,kt}}(x_{m,kt})$$

Since BP is used for $x_{m,kt}$, we have

$$b_{x_{m,kt}}(x_{m,kt}) = b_{\delta_{m,lt}; x_{m,kt}}(x_{m,kt})\tag{9.12}$$

In most cases, 4^M -QAM is used. To obtain the belief, we first look at the message $\delta_{m,lt} \rightarrow x_{m,kt}$,

$$\begin{aligned}\tau_{x_{m,kt}|\mathbf{y}} &= \left(\sum_l \tau_{\delta_{m,lt}; x_{m,kt}}^{-1} \right)^{-1} \\ m_{x_{m,kt}|\mathbf{y}} &= \tau_{x_{m,kt}|\mathbf{y}} \left(\sum_l \tau_{\delta_{m,lt}; x_{m,kt}}^{-1} m_{\delta_{m,lt}; x_{m,kt}} \right).\end{aligned}\tag{9.13}$$

Therefore,

$$b_{x_{m,kt}}(x_{m,kt}) = \frac{p(x_{m,kt}) \mathcal{CN}(x_{m,kt} | m_{x_{m,kt}|\mathbf{y}}, \tau_{x_{m,kt}|\mathbf{y}})}{Z_{x_{m,kt}}},\tag{9.14}$$

where

$$Z_{x_{m,kt}} = \sum_{x_{m,kt} \in S_m} p(x_{m,kt}) \mathcal{CN}(x_{m,kt} | m_{x_{m,kt}|\mathbf{y}}, \tau_{x_{m,kt}|\mathbf{y}}).\tag{9.15}$$

In the following, we denote the mean and variance of $b_{x_{m,kt}}$ as $m_{b_{x_{m,kt}}}$ and $\tau_{b_{x_{m,kt}}}$.

9.3.2 Message from $\delta_{m,lt}$ to $\mathbf{h}_{lk}$

With the CLT approximation, the joint belief pdf (9.8) can be written as

$$b_{\delta_{m,lt}}(\mathbf{h}_{lk}, x_{m,kt}) = \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{\mathbf{h}_{lk}|x_{m,kt}}, \mathbf{C}_{\mathbf{h}_{lk}|\sigma_{xm}^2}) b_{\delta_{m,lt};x_{m,kt}}(x_{m,kt}), \quad (9.16)$$

where

$$\begin{aligned} \mathbf{C}_{\mathbf{h}_{lk}|\sigma_{xm}^2} &= (\sigma_{xm}^2 \mathbf{C}_{\mathbf{u}_{m,lt}}^{-1} + \mathbf{C}_{\mathbf{h}_{lk};\delta_{m,lt}}^{-1})^{-1} \\ \mathbf{m}_{\mathbf{h}_{lk}|x_{m,kt}} &= \mathbf{C}_{\mathbf{h}_{lk}|\sigma_{xm}^2} \left(x_{m,kt}^* \mathbf{C}_{\mathbf{u}_{m,lt}}^{-1} \mathbf{m}_{\mathbf{u}_{m,lt}} + \mathbf{C}_{\mathbf{h}_{lk};\delta_{m,lt}}^{-1} \mathbf{m}_{\mathbf{h}_{lk};\delta_{m,lt}} \right) \end{aligned} \quad (9.17)$$

The mean and covariance matrix of belief $b_{\delta_{m,lt};\mathbf{h}_{lk}}$ for $\mathbf{h}_{lk}$ at $\delta_{m,lt}$ can be obtained as [56]:

$$\begin{aligned} \mathbf{m}_{b_{\delta_{m,lt};\mathbf{h}_{lk}}} &= \mathbf{C}_{\mathbf{h}_{lk}|\sigma_{xm}^2} \left(m_{b_{x_{m,kt}}}^* \mathbf{C}_{\mathbf{u}_{m,lt}}^{-1} \mathbf{m}_{\mathbf{u}_{m,lt}} + \mathbf{C}_{\mathbf{h}_{lk};\delta_{m,lt}}^{-1} \mathbf{m}_{\mathbf{h}_{lk};\delta_{m,lt}} \right) \\ \mathbf{C}_{b_{\delta_{m,lt};\mathbf{h}_{lk}}}^{-1} &= \mathbf{C}_{\mathbf{h}_{lk}|\sigma_{xm}^2}^{-1} - \tau_{b_{x_{m,kt}}} \mathbf{C}_{\mathbf{u}_{m,lt}}^{-1} \mathbf{m}_{\mathbf{u}_{m,lt}} \\ &\quad \cdot \left(1 + \tau_{b_{x_{m,kt}}} \mathbf{m}_{\mathbf{u}_{m,lt}}^H \mathbf{C}_{\mathbf{u}_{m,lt}}^{-1} \mathbf{C}_{\mathbf{h}_{lk}|\sigma_{xm}^2} \mathbf{C}_{\mathbf{u}_{m,lt}}^{-1} \mathbf{m}_{\mathbf{u}_{m,lt}} \right)^{-1} \mathbf{m}_{\mathbf{u}_{m,lt}}^H \mathbf{C}_{\mathbf{u}_{m,lt}}^{-1}. \end{aligned} \quad (9.18)$$

The feedback message can be obtained by $\frac{\text{proj}[b_{\delta_{m,lt};\mathbf{h}_{lk}}]}{\mu_{\mathbf{h}_{lk};\delta_{m,lt}}}$:

$$\begin{aligned} \mathbf{C}_{\delta_{m,lt};\mathbf{h}_{lk}} &= \frac{\mathbf{C}_{\mathbf{u}_{m,lt}}}{\sigma_{xm}^2} - \tau_{b_{x_{m,kt}}} \mathbf{m}_{\mathbf{u}_{m,lt}} \\ &\quad \cdot \left[\tau_{b_{x_{m,kt}}} \gamma_{aux_{m,lt}} - \sigma_{xm}^4 \right]^{-1} \mathbf{m}_{\mathbf{u}_{m,lt}}^H \\ \mathbf{m}_{\delta_{m,lt};\mathbf{h}_{lk}} &= \mathbf{C}_{\delta_{m,lt};\mathbf{h}_{lk}} \mathbf{C}_{b_{\delta_{m,lt};\mathbf{h}_{lk}}}^{-1} \mathbf{m}_{b_{\delta_{m,lt};\mathbf{h}_{lk}}} \\ &\quad - \mathbf{C}_{\mathbf{h}_{lk};\delta_{m,lt}}^{-1} \mathbf{m}_{\mathbf{h}_{lk};\delta_{m,lt}}, \end{aligned} \quad (9.19)$$

where

$$\gamma_{aux_{m,lt}} = \mathbf{m}_{\mathbf{u}_{m,lt}}^H \left(\frac{\mathbf{C}_{\mathbf{u}_{m,lt}}}{\sigma_{xm}^2} + \mathbf{C}_{\mathbf{h}_{lk};\delta_{m,lt}} \right)^{-1} \mathbf{m}_{\mathbf{u}_{m,lt}} \quad (9.20)$$

9.3.3 Message from $\delta_{m,lt}$ to $\mathbf{z}_{m,lt}$

Similar to the computation of $\mathbf{w}_{m,lt}$ in (9.9), we approximate the marginalization over $\forall k$, $\mathbf{h}_{lk}$ and $x_{m,kt}$ by CLT. Therefore, from (9.8), we have

$$\begin{aligned} b_{\delta_{m,lt}}(\mathbf{z}_{m,lt}) &= \mu_{\mathbf{z}_{m,lt};\delta_{m,lt}}(\mathbf{z}_{m,lt}) \\ &\quad \cdot \int \delta(\mathbf{z}_{m,lt} - \sum_k \mathbf{h}_{lk} x_{m,kt}) \prod_k \mu_{\mathbf{h}_{lk};\delta_{m,lt}}(\mathbf{h}_{lk}) \\ &\quad \cdot \prod_k \mu_{x_{m,kt};\delta_{m,lt}}(x_{m,kt}) d\mathbf{h}_{lk} dx_{m,kt} \end{aligned} \quad (9.21)$$

According to CLT, the integral in (9.21) can be approximated as Gaussian. Following feedback message rule, we can verify that the integral part is the feedback message, and thus,

$$\begin{aligned} \mathbf{m}_{\delta_{m,lt};\mathbf{z}_{m,lt}} &= \frac{1}{K-1} \sum_k \mathbf{m}_{\mathbf{w}_{m,l,k,t}} \\ \mathbf{C}_{\delta_{m,lt};\mathbf{z}_{m,lt}} &= \frac{1}{K-1} \sum_k \mathbf{C}_{\mathbf{w}_{m,l,k,t}}. \end{aligned} \quad (9.22)$$

9.3.4 Message from $f_{\mathbf{H}_{lG_g}}$ to $\mathbf{h}_{lk}$

Since $f_{\mathbf{H}_{lG_g}}$ is already Gaussian, the belief at $f_{\mathbf{H}_{lG_g}}$ must be Gaussian,

$$b_{f_{\mathbf{H}_{lG_g}}} = f_{\mathbf{H}_{lG_g}}(\mathbf{H}_{lG_g}) \prod_{k \in G_g} \mu_{\mathbf{h}_{lk}; f_{\mathbf{H}_{lG_g}}}(\mathbf{h}_{lk}). \quad (9.23)$$

Following [56], we get

$$\begin{aligned} & \mu_{f_{\mathbf{H}_{lG_g}}; \mathbf{h}_{lk}}(\mathbf{h}_{lk}) \\ &= \mathcal{CN} \left(\mathbf{h}_{lk} | \tilde{\mathbf{y}}_{p,lg} - \sum_{k' \in G_g / \{k\}} \mathbf{m}_{\mathbf{h}_{lk'} | \mathbf{y}}, \mathbf{C}_{\tilde{\mathbf{y}}} + \sum_{k' \in G_g / \{k\}} \mathbf{C}_{\mathbf{h}_{lk'} | \mathbf{y}} \right) \\ & \cdot \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{0}, \mathbf{\Xi}_{\mathbf{h}_{lk}}), \end{aligned} \quad (9.24)$$

where

$$\begin{aligned} \mathbf{C}_{\mathbf{h}_{lk} | \mathbf{y}} &= \left(\mathbf{C}_{\mathbf{h}_{lk}; f_{\mathbf{H}_{lG_g}}}^{-1} + \mathbf{\Xi}_{\mathbf{h}_{lk}}^{-1} \right)^{-1} \\ \mathbf{m}_{\mathbf{h}_{lk} | \mathbf{y}} &= \mathbf{C}_{\mathbf{h}_{lk} | \mathbf{y}} \mathbf{C}_{\mathbf{h}_{lk}; f_{\mathbf{H}_{lG_g}}}^{-1} \mathbf{m}_{\mathbf{h}_{lk}; f_{\mathbf{H}_{lG_g}}}. \end{aligned} \quad (9.25)$$

9.3.5 Message from $f_{\mathbf{y}_{lt}}$ to $\mathbf{z}_{m,lt}$

Since belief at $f_{\mathbf{y}_{lt}}$ is

$$b_{f_{\mathbf{y}_{lt}}} = p(\mathbf{y}_{lt} | \sum_m \mathbf{z}_{m,lt}) \prod_m \mu_{\mathbf{z}_{m,lt}; f_{\mathbf{y}_{lt}}}(\mathbf{z}_{m,lt}), \quad (9.26)$$

which is already a Gaussian, we can immediately derive the feedback message by Gaussian Reproduction Lemma [15]:

$$\begin{aligned} \mathbf{C}_{f_{\mathbf{y}_{lt}}; \mathbf{z}_{m,lt}} &= \mathbf{C}_v + \sum_{m' \neq m} \mathbf{C}_{\mathbf{z}_{m',lt}; f_{\mathbf{y}_{lt}}} \\ \mathbf{m}_{f_{\mathbf{y}_{lt}}; \mathbf{z}_{m,lt}} &= \mathbf{y}_{lt} - \sum_{m' \neq m} \mathbf{m}_{\mathbf{z}_{m',lt}; f_{\mathbf{y}_{lt}}} \end{aligned} \quad (9.27)$$

9.4 Simulation Results

We simulate an environment within a 400×400 square meter area, equipped with 16 APs and 8 User Terminals. Each AP features $N = 2$ antennas and is positioned at coordinates $(\frac{400}{3}i, \frac{400}{3}j)$, $i, j \in \{0, 1, 2, 3\}$. The UTs are uniformly distributed throughout the area. We denote the distance between each UT k and AP l as d_{lk} . Channel covariances for each user k at AP l are modeled using $N \times N$ diagonal matrices, represented as $\sigma_{h_{lk}}^2 \mathbf{I}$, where $10 \log_{10}(\sigma_{h_{lk}}^2) = -30 - 36.7 \log_{10}(d_{lk})$.

The length of the orthogonal pilot sequences is set to $P = 6$ to introduce pilot contamination.

We employ 4^M QAM constellation of length $T = 16$ for signal transmission where $M \in \{1, 2\}$. Following [57], we employed power control to ensure the total received power at the APs is the same. We base our results on 50 different realizations, which are illustrated in Figure 9.2. In the Genie-Aided scenario, we implement the proposed algorithm as if the data symbols are known. In the MMSE Genie-Aided scenario, all the APs estimate the channel coefficients using the MMSE estimator with known data sequences, which is the theoretical lower bound for all channel estimation methods.

Algorithm 9.1: One Iteration of Decentralized EP

Require: $\Xi_{h_{lk}}, \tilde{\mathbf{y}}_{p,lg}, \mathbf{y}_{lt}, \sigma_x^2, \sigma_v^2, G_g, \mathcal{S}$

- 1: Initialize $\mu_{\delta_{m,lt};h_{lk}}, \mu_{\delta_{m,lt};z_{m,lt}}, \mu_{\delta_{m,lt};x_{m,kt}},$
 - 2: $\mathbf{C}_{f_{H_{lg}};h_{lk}}$ and $\mathbf{m}_{f_{H_{lg}};h_{lk}}$ via (9.25) $\rightarrow$ (9.24)
 - 3: $\mathbf{C}_{f_{y_{lt}};z_{m,lt}}$ and $\mathbf{m}_{f_{y_{lt}};z_{m,lt}}$ via (9.27)
 - 4: $\mathbf{C}_{u_{m,lt}}^{\mathbf{u}_{m,lt}}$ and $\mathbf{m}_{u_{m,lt}}$ via (9.10)
 - 5: $\tau_{\delta_{m,lt};x_{m,kt}}$ and $m_{\delta_{m,lt};x_{m,kt}}$ via (9.11)
 - 6: $b_{x_{m,kt}}(x_{m,kt})$ via (9.13) $\rightarrow$ (9.15) $\rightarrow$ (9.14)
 - 7: $\mathbf{C}_{h_{lk}|\sigma_x^2}$ via (9.17)
 - 8: $\mathbf{C}_{b_{\delta_{m,lt};h_{lk}}}^{-1}$ and $\mathbf{m}_{b_{\delta_{m,lt};h_{lk}}}$ via (9.18)
 - 9: $\gamma_{au_{m,lt}}$ via (9.20)
 - 10: $\mathbf{C}_{\delta_{m,lt};h_{lk}}^{-1} \mathbf{m}_{\delta_{m,lt};h_{lk}}$ and $\mathbf{P}_{\delta_{m,lt};h_{lk}}$ via (9.19)
 - 11: $\mathbf{C}_{\delta_{m,lt};z_{m,lt}}$ and $\mathbf{m}_{\delta_{m,lt};z_{m,lt}}$ via (9.22)
-

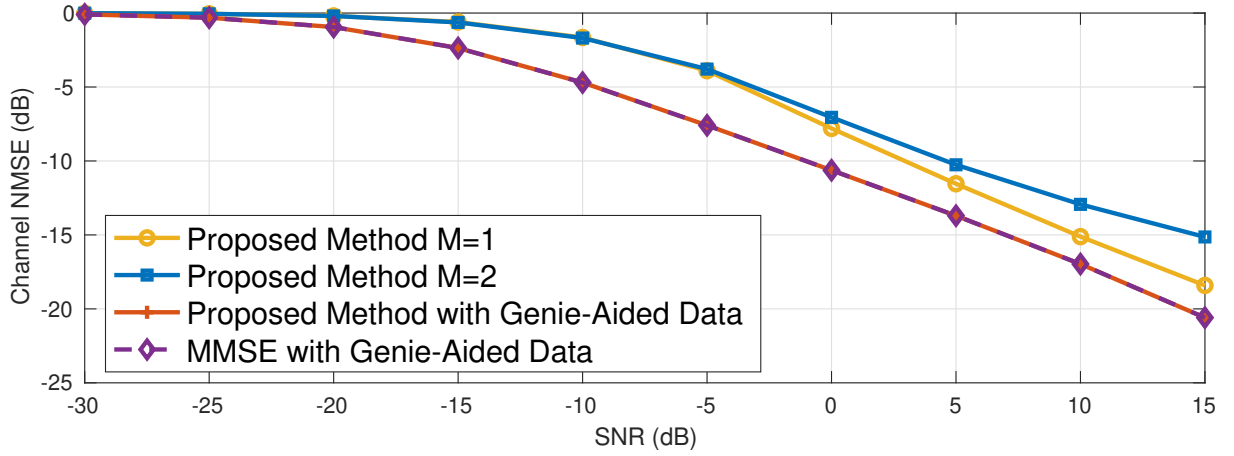


Figure 9.2: NMSE vs SNR

9.5 Conclusions

This chapter introduces a simplified, distributed EP-BP-based algorithm for bilinear joint estimation. By leveraging the structure of the data symbols, we reduce the complexity associated with the constellation size. The proposed algorithm is compatible with 4^M -QAM data symbols. At each AP, the computational complexity is $O(MKT \cdot N^3)$, while the CPU handles the computation of data symbol beliefs with a complexity of $O(MLKT)$.

This approach can be extended to other message-passing algorithms to further reduce complexity. For example, the hybrid VB-EP algorithm in [61] treats the entire data sequence as an atomic variable. In contrast, our method considers each individual data symbol as an atomic variable.

Chapter 10

Distributed Iterative ML and Message Passing for Grant-Free Cell-Free Massive MIMO Systems

10.1 Introduction

Cell-Free (CF) Massive Multiple-Input Multiple-Output (MaMIMO) has attracted significant attention due to its ability to leverage the benefits of both MaMIMO and distributed antenna systems [1, 20]. In CF MaMIMO systems, a large number of Access Points (APs) are geographically distributed over a wide area. Although a Central Processing Unit (CPU) coordinates the system, the information exchange between APs and the CPU is limited. Consequently, APs are unable to share their Channel State Information (CSI), which necessitates the development of distributed algorithms for data detection and channel estimation [1, 2, 62].

With the sporadic traffic patterns introduced by the Internet of Things (IoT), the Grant-Free (GF) access scheme has been proposed as a reliable, low-latency solution [4]. Moreover, it has been shown that MaMIMO systems are well-suited to support GF access.

Due to the large coverage area of CF MaMIMO systems and the absence of cellular boundaries, the number of User Terminals (UTs) often exceeds the number of available orthogonal pilot sequences. This results in the well-known issue of pilot contamination, which prevents the accurate acquisition of Channel State Information (CSI) [20]. To address this challenge, semi-blind approaches—where Access Points (APs) jointly estimate the channel and detect user symbols—have been investigated in [51]. However, the joint estimation of CSI and data leads to an intractable high-dimensional inference problem, necessitating the use of approximate inference techniques.

10.1.1 Prior Works

In [61], a hybrid Variational Bayes and Expectation Propagation (VB-EP) algorithm was proposed for semi-blind estimation by minimizing the Bethe Free Energy (BFE) [6]. However, this method is sensitive to initialization and may diverge away from the stable point with bad initializations, particularly when user data are drawn from a finite alphabet. Moreover, it is not suitable for GF access schemes.

The Variable-Level EP (VL-EP) algorithm [20] offers improved robustness when user data follow a Gaussian distribution. It can be interpreted as an approximate maximum a pos-

teriori (MAP) estimator for the channel coefficients, due to its Expectation-Maximization (EM) update mechanism. However, VL-EP still suffers from divergence when finite alphabet data are used. Another limitation is that only a centralized version of this algorithm was proposed in [20]. In contrast to VB-EP, VL-EP treats multiuser interference as zero-mean noise when estimating a user's channel based on the data likelihood, which results in performance saturation in the high signal-to-noise ratio (SNR) regime.

To improve convergence, the concept of introducing a pseudo prior into message-passing (MP)-based algorithms has been proposed in [63]. A pseudo prior is essentially an approximate posterior, typically obtained from a simpler algorithm. In MP algorithms, it is incorporated by multiplying the pseudo prior with the true prior (or 1, if the parameter is deterministic) to form an enhanced prior. From this perspective, the pseudo prior does not introduce new information; rather, it serves as a regularization term that helps prevent the algorithm from diverging.

10.1.2 Main Contributions

In this chapter, we propose a distributed algorithm for activity detection, channel estimation, and data detection. We observe that when pilot sequences are mutually orthogonal and the data symbols have zero mean, the Central Limit Theorem (CLT) can be applied to enable an approximate ML algorithm for distributed activity detection. After the activity detection, we treat the activity indicator as given and perform the bilinear channel estimation and data detection jointly. During the bilinear estimation, we leverage the concept of a pseudo prior to combine the strengths of VL-EP and VB-EP. We use the results in VL-EP as pseudo prior for VB-EP to prevent VB-EP from drifting away from the critical point.

10.2 System Model

We consider a semi-blind uplink signal model comprising L APs and $\bar{K}$ single-antenna UTs. Each AP contains N antennas. The local system model at the l -th AP is given by

$$\begin{bmatrix} \mathbf{Y}_{p,l} & \mathbf{Y}_l \end{bmatrix} = \bar{\mathbf{H}}_l \mathbf{U} \begin{bmatrix} \bar{\mathbf{X}}_p & \bar{\mathbf{X}} \end{bmatrix} + \begin{bmatrix} \mathbf{V}_{p,l} & \mathbf{V}_l \end{bmatrix}. \quad (10.1)$$

The received signals at the l -th AP consists of a pilot part $\mathbf{Y}_{p,l} \in \mathbb{C}^{N \times P}$ and a data part $\mathbf{Y}_l \in \mathbb{C}^{N \times T}$. Each column $\bar{\mathbf{h}}_{l\bar{k}}$ of $\bar{\mathbf{H}}_l$ represents the uplink channel from UT $\bar{k}$ to AP l . The channel $\bar{\mathbf{h}}_{l\bar{k}}$ is modeled as a zero mean Gaussian random vector with covariance matrix $\bar{\boldsymbol{\Sigma}}_{\bar{\mathbf{h}}_{l\bar{k}}} \in \mathbb{C}^{N \times N}$. The activity indicator matrix is denoted by $\mathbf{U} = \text{diag}[u_1, \dots, u_{\bar{K}}]$, which is an unknown deterministic diagonal matrix with binary entries. Its diagonal elements form a vector $\mathbf{u} = \text{vec}(\mathbf{U})$. The transmitted signal matrix is composed of pilot symbols $\bar{\mathbf{X}}_p \in \mathbb{C}^{\bar{K} \times P}$ and data symbols $\bar{\mathbf{X}} \in \mathcal{S}^{\bar{K} \times T}$, where $\mathcal{S}$ denotes the constellation set. We denote $\bar{\mathbf{x}}_{\bar{k}}^T$ as the $\bar{k}$ -th row of $\bar{\mathbf{X}}$, and it represents the data sequence transmitted by user $\bar{k}$. The signal power is denoted as σ_x^2 . The noise is considered as i.i.d. Gaussian distribution, and thus, $\text{vec}(\begin{bmatrix} \mathbf{V}_{p,l} & \mathbf{V}_l \end{bmatrix}) \sim \mathcal{CN}(0, \mathbf{C}_v)$ with $\mathbf{C}_v = \sigma_v^2 \mathbf{I}$. We denote $\mathbf{Y}_p$, $\mathbf{Y}$, $\bar{\mathbf{H}}$, $\mathbf{V}_p$ and $\mathbf{V}$ as global system parameter.

10.2.1 Orthogonal Pilot sequences

If orthogonal pilot sequences are used, we can first preprocess the pilot observation by right multiplying it with $\tilde{\mathbf{x}}_{p,g}^*$ which is the conjugated g -th pilot sequence from the pilot

codebook. This results in an equivalent observation $\tilde{\mathbf{y}}_{p,lg}$

$$\tilde{\mathbf{y}}_{p,lg} = \frac{1}{P\sigma_x^2} \mathbf{Y}_{p,l} \tilde{\mathbf{x}}_{p,g}^* = \sum_{\bar{k} \in \bar{G}_g} \bar{\mathbf{h}}_{l\bar{k}} u_{\bar{k}} + \tilde{\mathbf{v}}_{p,lg} \quad (10.2)$$

where $\tilde{\mathbf{v}}_{p,lg} = \frac{1}{P\sigma_x^2} \mathbf{V}_{p,l} \tilde{\mathbf{x}}_{p,g}^* \sim \mathcal{N}(\mathbf{v}_{p,lg} | \mathbf{0}, \mathbf{C}_{\tilde{\mathbf{v}}})$, $\mathbf{C}_{\tilde{\mathbf{v}}} = \frac{\sigma_v^2 \mathbf{I}}{P\sigma_x^2}$, and $\bar{G}_g$ denotes the set of users using the g -th pilot sequence. We observe that every $\bar{\mathbf{h}}_{l\bar{k}} u_{\bar{k}}$ occurs only in one group $\bar{G}_g$, and the cross-correlation $\mathbb{E}[\tilde{\mathbf{v}}_{p,lg} \tilde{\mathbf{v}}_{p,lg'}^H]$ is an all-zero matrix for all $g \neq g'$. Therefore, the observations $\tilde{\mathbf{y}}_{p,lg}$ and $\tilde{\mathbf{y}}_{p,lg'}$ are independent. Denote $\bar{\mathbf{H}}_{l\bar{G}_g}$ as the matrix stacking $\forall \bar{k} \in \bar{G}_g, \bar{\mathbf{h}}_{l\bar{k}}$ as columns.

10.3 Iterative ML for User Activity Detection

The method we use for activity detection is based on maximum likelihood. Denote the estimated activity indicator as $\hat{\mathbf{u}} = \text{Diag}[\hat{\mathbf{U}}]$. The likelihood of $\hat{\mathbf{u}}$ is

$$p(\mathbf{Y}, \mathbf{Y}_p | \hat{\mathbf{U}}) = \int p(\mathbf{Y} | \bar{\mathbf{H}}, \hat{\mathbf{U}}, \bar{\mathbf{X}}) p(\bar{\mathbf{X}}) \cdot \prod_l p(\bar{\mathbf{H}}_l) \prod_g p(\tilde{\mathbf{y}}_{p,lg} | \bar{\mathbf{H}}_{l\bar{G}_g}, \hat{\mathbf{U}}) d\bar{\mathbf{H}} d\bar{\mathbf{X}} \quad (10.3)$$

Before investigating (10.3), we introduce a useful relation

$$\begin{aligned} \mathcal{CN}(\mathbf{m}_1 | \mathbf{A}\mathbf{x}, \mathbf{C}_1) \mathcal{CN}(\mathbf{x} | \mathbf{m}_2, \mathbf{C}_2) \\ = \mathcal{CN}(\mathbf{x} | \mathbf{m}_3, \mathbf{C}_3) \mathcal{CN}(\mathbf{m}_1 | \mathbf{A}\mathbf{m}_2, \mathbf{C}_1 + \mathbf{A}\mathbf{C}_2\mathbf{A}^H) \end{aligned} \quad (10.4)$$

where

$$\begin{aligned} \mathbf{C}_3 &= (\mathbf{A}^H \mathbf{C}_1^{-1} \mathbf{A} + \mathbf{C}_2^{-1})^{-1} \\ \mathbf{m}_3 &= \mathbf{C}_3 (\mathbf{A}^H \mathbf{C}_1^{-1} \mathbf{m}_1 + \mathbf{C}_2^{-1} \mathbf{m}_2). \end{aligned} \quad (10.5)$$

This relation can be verified by noticing

$$\begin{aligned} \mathbf{D}(\mathbf{A} + \mathbf{BCD})^{-1} &= \mathbf{C}^{-1}(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1})^{-1} \mathbf{DA}^{-1}; \\ \det(\mathbf{A} + \mathbf{BCD}) &= \det(\mathbf{A}) \det(\mathbf{B}) \det(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B}). \end{aligned}$$

We combine pilot observation and channel prior to get a pilot based prior $\tilde{p}(\bar{\mathbf{H}}_{l\bar{G}_g} | \hat{\mathbf{u}})$ (even though it is a posterior given $\mathbf{Y}_p$, we can still understand it as prior information as if we have a blind estimation problem)

$$\begin{aligned} \tilde{p}(\bar{\mathbf{H}}_{l\bar{G}_g} | \hat{\mathbf{u}}) &= p(\bar{\mathbf{H}}_{l\bar{G}_g} | \mathbf{Y}_p, \hat{\mathbf{u}}) \propto p(\tilde{\mathbf{y}}_{p,lg}, \bar{\mathbf{h}}_{l\bar{G}_g} | \hat{\mathbf{u}}) \\ &= \mathcal{CN}(\tilde{\mathbf{y}}_{p,lg} | \sum_{\bar{k} \in \bar{G}_g \cap U(\hat{\mathbf{u}})} \bar{\mathbf{h}}_{l\bar{k}}, \mathbf{C}_{\tilde{\mathbf{v}}}) \mathcal{CN}(\bar{\mathbf{h}}_{l\bar{G}_g} | \mathbf{0}, \mathbf{\Xi}_{\bar{\mathbf{h}}_{l\bar{G}_g}}) \end{aligned} \quad (10.6)$$

Assume $\bar{k} \in \bar{G}_g$, and in later context, we will use the marginal pilot-based prior $\tilde{p}(\bar{\mathbf{h}}_{l\bar{k}})$ which is computed from

$$\int p(\tilde{\mathbf{y}}_{p,lg}, \bar{\mathbf{h}}_{l\bar{G}_g} | \hat{\mathbf{u}}) d\bar{\mathbf{h}}_{l\{\bar{k}' \neq \bar{k}\}} = \mathcal{CN}(\tilde{\mathbf{y}}_{p,lg} | \mathbf{0}, \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg} | \hat{\mathbf{u}}}) \tilde{p}(\bar{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}})$$

where

$$\begin{aligned} \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg}|\hat{\mathbf{u}}} &= \mathbf{C}_{\tilde{\mathbf{v}}} + \sum_{\bar{k} \in \bar{G}_g \cap U(\hat{\mathbf{u}})} \boldsymbol{\Xi}_{\bar{\mathbf{h}}_{l\bar{k}}}; \\ \tilde{p}(\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}) &= p(\hat{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}, \mathbf{Y}_p) = \mathcal{CN}(\bar{\mathbf{h}}_{l\bar{k}}|\mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}}, \mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}}) \end{aligned} \quad (10.7)$$

and

$$\begin{aligned} \mathbf{C}_{\mathbf{Y}_p|\bar{\mathbf{h}}_{l\bar{k}},\hat{\mathbf{u}}} &= \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg}|\hat{\mathbf{u}}} - \boldsymbol{\Xi}_{\bar{\mathbf{h}}_{l\bar{k}}}; \\ \mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} &= (\mathbf{C}_{\mathbf{Y}_p|\bar{\mathbf{h}}_{l\bar{k}},\hat{\mathbf{u}}} + \boldsymbol{\Xi}_{\bar{\mathbf{h}}_{l\bar{k}}}^{-1})^{-1} \\ \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} &= \mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{C}_{\mathbf{Y}_p|\bar{\mathbf{h}}_{l\bar{k}},\hat{\mathbf{u}}}^{-1} \tilde{\mathbf{y}}_{p,lg}. \end{aligned} \quad (10.8)$$

Now we take data observation into consideration. Rewrite (10.3),

$$\begin{aligned} p(\mathbf{Y}, \mathbf{Y}_p|\hat{\mathbf{U}}) &= \left[\prod_{l,g} \mathcal{CN}(\tilde{\mathbf{y}}_{p,lg}|\mathbf{0}, \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg}|\hat{\mathbf{u}}}) \right] \\ &\cdot \int p(\mathbf{Y}|\sum_{\bar{k} \in U(\hat{\mathbf{u}})} \bar{\mathbf{x}}_{\bar{k}} \otimes \bar{\mathbf{h}}_{\bar{k}}) p(\bar{\mathbf{X}}) \cdot \prod_{l,g} \tilde{p}(\bar{\mathbf{H}}_{l\bar{G}_g}|\hat{\mathbf{U}}) d\mathbf{H} d\mathbf{X} \end{aligned} \quad (10.9)$$

where we use CLT to approximate the summation. Suppose that $\bar{k} \in U(\hat{\mathbf{u}})$, and we stack the results from (10.8) to form

$$\begin{aligned} \mathbf{C}_{\bar{\mathbf{h}}_{\bar{k}}|\hat{\mathbf{u}}} &:= \text{cov}[\bar{\mathbf{h}}_{\bar{k}}] \Rightarrow [\mathbf{C}_{\bar{\mathbf{h}}_{\bar{k}}|\hat{\mathbf{u}}}]_{ll'}^{(N)} = \delta_{ll'} \mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} \\ \mathbf{m}_{\bar{\mathbf{h}}_{\bar{k}}|\hat{\mathbf{u}}} &:= \mathbb{E}[\bar{\mathbf{h}}_{\bar{k}}] \Rightarrow [\mathbf{m}_{\bar{\mathbf{h}}_{\bar{k}}|\hat{\mathbf{u}}}]_l^{(N)} = \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} \end{aligned} \quad (10.10)$$

The cross correlation between different $\bar{\mathbf{z}}_{\bar{k}}$ is not important, since $\forall \bar{k} \neq \bar{k}', \mathbb{E}[\bar{\mathbf{x}}_{\bar{k}} \bar{\mathbf{x}}_{\bar{k}'}^H] = \mathbf{0}$. Denote $\bar{\mathbf{z}}_{\bar{k}} = \bar{\mathbf{x}}_{\bar{k}} \otimes \bar{\mathbf{h}}_{\bar{k}}$ which has zero mean and covariance

$$\mathbf{C}_{\bar{\mathbf{z}}_{\bar{k}}} = \mathbb{E}[\bar{\mathbf{z}}_{\bar{k}} \bar{\mathbf{z}}_{\bar{k}}^H] = \sigma_x^2 [\mathbf{I}_T \otimes (\mathbf{C}_{\bar{\mathbf{h}}_{\bar{k}}|\hat{\mathbf{u}}} + \mathbf{m}_{\bar{\mathbf{h}}_{\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{\mathbf{h}}_{\bar{k}}|\hat{\mathbf{u}}}^H)] \quad (10.11)$$

Therefore, the likelihood can be represented by

$$\begin{aligned} p(\mathbf{Y}, \mathbf{Y}_p|\hat{\mathbf{U}}) &= p(\mathbf{Y}|\mathbf{Y}_p, \hat{\mathbf{U}}) p(\mathbf{Y}_p|\hat{\mathbf{U}}) \\ &\simeq \mathcal{CN}(\mathbf{y}|\mathbf{0}, \mathbf{C}_{\mathbf{Y}|\mathbf{Y}_p,\hat{\mathbf{u}}}) \prod_{l,g} \mathcal{CN}(\tilde{\mathbf{y}}_{p,lg}|\mathbf{0}, \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg}|\hat{\mathbf{u}}}) \end{aligned} \quad (10.12)$$

where

$$\mathbf{C}_{\mathbf{Y}|\mathbf{Y}_p,\hat{\mathbf{u}}} = \mathbf{C}_v + \sum_{\bar{k} \in U(\hat{\mathbf{u}})} \mathbf{C}_{\bar{\mathbf{z}}_{\bar{k}}}. \quad (10.13)$$

The log-likelihood of $\hat{\mathbf{u}}$ is $\ln p(\mathbf{Y}, \mathbf{Y}_p|\hat{\mathbf{U}}) \simeq L_p(\hat{\mathbf{u}}) + L_d(\hat{\mathbf{u}})$, where

$$\begin{aligned} L_p(\hat{\mathbf{u}}) &= \sum_{l,g} \ln \mathcal{CN}(\tilde{\mathbf{y}}_{p,lg}|\mathbf{0}, \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg}|\hat{\mathbf{u}}}) \\ &= \sum_{l,g} -\ln[\det(\mathbf{C}_{\tilde{\mathbf{y}}_{p,lg}|\hat{\mathbf{u}}})] - \tilde{\mathbf{y}}_{p,lg}^H \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg}|\hat{\mathbf{u}}}^{-1} \tilde{\mathbf{y}}_{p,lg} \\ L_d(\hat{\mathbf{u}}) &= \ln \mathcal{CN}(\mathbf{y}|\mathbf{0}, \mathbf{C}_{\mathbf{Y}|\mathbf{Y}_p,\hat{\mathbf{u}}}) \\ &= -\ln[\det(\mathbf{C}_{\mathbf{Y}|\mathbf{Y}_p,\hat{\mathbf{u}}})] - \mathbf{y}^H \mathbf{C}_{\mathbf{Y}|\mathbf{Y}_p,\hat{\mathbf{u}}}^{-1} \mathbf{y}. \end{aligned} \quad (10.14)$$

Algorithm 10.1: Iterative ML

- 1: Initialize: $i = 1, \hat{\mathbf{u}} = \mathbf{0}$
 - 2: **repeat**
 - 3: $\bar{k} = [(i - 1) \bmod \bar{K}] + 1$
 - 4: $\hat{u}_k = \arg \max_{\theta \in \{0,1\}} L_{tot}((\mathbf{I} - \mathbf{E}_{\bar{k}})\hat{\mathbf{u}} + \mathbf{e}_{\bar{k}}\theta)$
 - 5: **until** Convergence
-

Now we try to simplify L_d . Consider the large system limit, i.e., $L, \bar{K}, P, T \rightarrow \infty$ with fixed ratio between every pair. To have a realistic system, $[\mathbf{Y}]_{n,t}$ and $\mathbb{E}\{||[\bar{\mathbf{X}}]_{\bar{k},t}||^2\}$ scale with $O(1)$. Therefore, $\text{tr}[\Xi_{\bar{h}_{l\bar{k}}}]$ scales with $O(1/K)$. This indicates that the off diagonal blocks of $\mathbf{C}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}} + \mathbf{m}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}}^H$ are higher-order perturbations of the blocks on the diagonal. We neglect the off diagonal blocks by approximating

$$\forall l \neq l'. \sum_{\bar{k} \in U(\hat{\mathbf{u}})} \mathbf{m}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{h}_{l'\bar{k}}|\hat{\mathbf{u}}}^H = \mathbf{0} \quad (10.15)$$

After that, $\mathbf{C}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}} + \mathbf{m}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}}^H$ is approximated as a block diagonal matrix with block size $N \times N$. This entails $\hat{L}_d \simeq L_d$ where

$$\begin{aligned} \hat{L}_d(\hat{\mathbf{u}}) \simeq & \sum_l \left\{ -T \ln \left(\det \left[\mathbf{C}_v + \sum_{\bar{k} \in U(\hat{\mathbf{u}})} \sigma_x^2 (\mathbf{C}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}} + \mathbf{m}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}}^H) \right] \right) \right. \\ & \left. - \sum_t \mathbf{y}_{lt}^H [\mathbf{C}_v + \sum_{\bar{k} \in U(\hat{\mathbf{u}})} \sigma_x^2 (\mathbf{C}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}} + \mathbf{m}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{h}_{l\bar{k}}|\hat{\mathbf{u}}}^H)]^{-1} \mathbf{y}_{lt} \right\} \end{aligned} \quad (10.16)$$

where $\mathbf{y}_{lt}$ is the t -th column of $\mathbf{Y}_l$. After the approximation, compared to $L_d(\hat{\mathbf{u}})$, $\hat{L}_d(\hat{\mathbf{u}})$ not only lowers the complexity, but also allows a distributed computation.

Combine both L_p and approximated $\hat{L}_d$ and the likelihood $\hat{\mathbf{u}}$ can be finally approximated as

$$L_{tot}(\hat{\mathbf{u}}) = \ln p(\mathbf{Y}, \mathbf{Y}_p | \hat{\mathbf{U}}) \simeq L_p(\hat{\mathbf{u}}) + \hat{L}_d(\hat{\mathbf{u}}). \quad (10.17)$$

The constrained ML is NP-hard. Therefore, we present here an iterative method as a compromise. During the i -th iteration, we update the $u_{\bar{k}}$ where $\bar{k} = [(i - 1) \bmod \bar{K}] + 1$ to maximize (10.17). We repeat this procedure until convergence. The complete procedure is outlined in Algorithm 10.1, where $\mathbf{e}_{\bar{k}}$ denotes a vector with a one in the $\bar{k}$ -th position and zeros elsewhere, and $\mathbf{E}_{\bar{k}} = \text{Diag}(\mathbf{e}_{\bar{k}})$ is its corresponding diagonal matrix.

10.4 Distributed VL-EP for Pseudo-Prior

With estimated $\hat{\mathbf{u}}$ from previous section, we prune the estimated non-active users

$$\mathbf{H} = \bar{\mathbf{H}}\hat{\mathbf{U}}; \mathbf{H}_l = \bar{\mathbf{H}}_l\hat{\mathbf{U}}; \mathbf{X} = \hat{\mathbf{U}}\bar{\mathbf{X}}; K = \hat{\mathbf{u}}^H \hat{\mathbf{u}} \quad (10.18)$$

with $\mathbf{h}_k, \mathbf{h}_{lk}, \mathbf{x}_k^T$ as the k -th column of $\mathbf{H}, \mathbf{H}_l$ and the k -th row of $\mathbf{X}$ respectively. Furthermore, we use $\Xi_{\mathbf{h}_{lk}}, G_g$ to denote a pruned version of $\Xi_{\bar{h}_{l\bar{k}}}$ and $\bar{G}_g$.

Distributed VL-EP can be used for pseudo-prior of VB-EP. To estimate the k -th user channel, we consider other channels as interference and alternating estimate all the channels, i.e., minimize the negative log pdf $-\ln p(\mathbf{Y}, \tilde{\mathbf{y}}_{p,lg}, \mathbf{h}_k)$ while treat $\forall k' \neq k. \mathbf{h}_{k'}$ as

interference with distribution $\mathcal{CN}(\mathbf{h}'_k | \mathbf{m}_{\check{\mathbf{h}}_{k'}}, \mathbf{C}_{\check{\mathbf{h}}_{k'}})$. At each AP, the noise plus interference of received data $\mathbf{y}_l = \text{vec}(\mathbf{Y}_l)$ has zero mean and covariance matrix

$$\mathbf{C}_{\check{\mathbf{v}}_{lk}} = \sigma_v^2 \mathbf{I}_{NT} + \sum_{k' \neq k} \sigma_x^2 \mathbf{I}_T \otimes (\mathbf{C}_{\check{\mathbf{h}}_{lk'}} + \mathbf{m}_{\check{\mathbf{h}}_{lk}} \mathbf{m}_{\check{\mathbf{h}}_{lk}}^H) = \mathbf{I}_T \otimes \mathbf{C}_{\check{\mathbf{v}}_{lkt}}$$

where

$$\mathbf{C}_{\check{\mathbf{v}}_{lkt}} = \sigma_v^2 \mathbf{I}_N + \sum_{k' \neq k} \sigma_x^2 (\mathbf{C}_{\check{\mathbf{h}}_{lk'}} + \mathbf{m}_{\check{\mathbf{h}}_{lk}} \mathbf{m}_{\check{\mathbf{h}}_{lk}}^H) \quad (10.19)$$

is constant in time since the noise power is constant in time and we define $\mathbf{C}_{\check{\mathbf{v}}_{lk0}} = \mathbf{C}_{\check{\mathbf{v}}_{lkt}}$. On the other hand, if we look at the pilot part. The noise plus interference of the received pilot at each AP $\tilde{\mathbf{y}}_{p,lg}$ relative to user k can be modeled as Gaussian with the following mean and covariance matrix

$$\mathbf{m}_{\check{\mathbf{v}}_{p,lk}} = \sum_{k' \in G_g / \{k\}} \mathbf{m}_{\check{\mathbf{h}}_{lk'}}; \quad \mathbf{C}_{\check{\mathbf{v}}_{p,lk}} = \mathbf{C}_{\check{\mathbf{v}}} + \sum_{k' \in G_g / \{k\}} \mathbf{C}_{\check{\mathbf{h}}_{lk'}} \quad (10.20)$$

In order to construct a majorization function, we first look at the approximated marginal joint pdf (with interference approximated as Gaussian noise)

$$\begin{aligned} p(\mathbf{Y}, \tilde{\mathbf{y}}_{p,lg}, \mathbf{h}_k, \mathbf{x}_k) &= p(\mathbf{Y} | \mathbf{h}_k, \mathbf{x}_k) p(\tilde{\mathbf{y}}_{p,lg} | \mathbf{h}_k) p(\mathbf{h}_k) p(\mathbf{x}_k) \\ &= p(\mathbf{x}_k) \prod_l p(\mathbf{Y}_l | \mathbf{h}_{lk}, \mathbf{x}_k) p(\tilde{\mathbf{y}}_{p,lg} | \mathbf{h}_{lk}) p(\mathbf{h}_{lk}) \end{aligned} \quad (10.21)$$

The effect term of the majorization function at $\mathbf{h}_k = \check{\mathbf{h}}_k$ is [20]

$$\text{em}(\mathbf{h}_k | \check{\mathbf{H}}) = -\mathbb{E}_{\mathbf{x}_k | \mathbf{Y}, \mathbf{m}_{\check{\mathbf{h}}_k}} [\ln p(\mathbf{Y}, \tilde{\mathbf{y}}_{p,lg}, \mathbf{h}_k, \mathbf{x}_k)]. \quad (10.22)$$

The channel $\mathbf{h}_k$ is estimated by $\arg \min_{\mathbf{h}_k} \text{em}(\mathbf{h}_k | \check{\mathbf{H}})$ iteratively.

10.4.1 Expectation Step

To minimize the majorization function (10.22), we need to get the posterior $p(\mathbf{x}_k | \mathbf{Y}, \mathbf{m}_{\check{\mathbf{h}}_k})$. Rewrite it into the product of prior and likelihood

$$p(\mathbf{x}_k | \mathbf{Y}, \mathbf{m}_{\check{\mathbf{h}}_k}) \propto p(\mathbf{x}_k) \prod_l p(\mathbf{Y}_l | \mathbf{x}_k, \mathbf{m}_{\check{\mathbf{h}}_{lk}}) \quad (10.23)$$

We observe that the likelihood of the observation at the l -th AP, which is proportional to Gaussian distribution:

$$p(\mathbf{Y}_l | \mathbf{x}_k, \mathbf{m}_{\check{\mathbf{h}}_{lk}}) \propto \mathcal{CN}(\mathbf{x}_k | \mathbf{m}_{\dot{\mathbf{x}}_{lk}}, \mathbf{C}_{\dot{\mathbf{x}}_{lk}}), \quad (10.24)$$

where

$$\begin{aligned} \mathbf{C}_{\dot{\mathbf{x}}_{lk}} &= \tau_{\dot{\mathbf{x}}_{lk}}^i \mathbf{I}_T; \quad m_{\dot{\mathbf{x}}_{lkt}} = \tau_{\dot{\mathbf{x}}_{lk}}^i \mathbf{m}_{\check{\mathbf{h}}_{lk}}^H \mathbf{C}_{\check{\mathbf{v}}_{lk0}}^{-1} \mathbf{y}_{lt}; \\ \tau_{\dot{\mathbf{x}}_{lk}} &= \left(\mathbf{m}_{\check{\mathbf{h}}_{lk}}^{(i-1)H} \mathbf{C}_{\check{\mathbf{v}}_{lk}}^{-1} \mathbf{m}_{\check{\mathbf{h}}_{lk}}^{(i-1)} \right)^{-1} \end{aligned} \quad (10.25)$$

Since $\mathbf{C}_{\dot{\mathbf{x}}_{lk}}$ is a diagonal matrix, we can compute the posterior statistics of x_{kt} by one-dimensional integrals. We denote marginalized version of (10.23)

$$\begin{aligned} \check{p}(x_{kt}) &= \int p(\mathbf{x}_k | \mathbf{Y}_l, \mathbf{m}_{\check{\mathbf{h}}_{lk}}) d\mathbf{x}_{\{k' \neq k\} \{t' \neq t\}} \\ &= p(x_{kt}) \prod_l \mathcal{CN}(x_{kt} | m_{\dot{\mathbf{x}}_{lkt}}, \tau_{\dot{\mathbf{x}}_{lk}}) \end{aligned} \quad (10.26)$$

Algorithm 10.2: VL-EP For Pseudo Prior

- 1: Initialize: $\tau_{\tilde{x}_{kt}} = 1, m_{\tilde{x}_{kt}} = 0$
 - 2: **repeat**
 - 3: $\mathbf{C}_{\tilde{\mathbf{h}}_{lk}}, \mathbf{m}_{\tilde{\mathbf{h}}_{lk}}$ by (10.29)
 - 4: $\mathbf{C}_{\tilde{\mathbf{v}}_{lk0}}, \mathbf{m}_{\tilde{\mathbf{v}}_{p,lk}}, \mathbf{C}_{\tilde{\mathbf{v}}_{p,lk}}$ by (10.19), (10.20)
 - 5: $\tau_{\tilde{x}_{kt}}, m_{\tilde{x}_{kt}}$ by (10.25), (10.27)
 - 6: **until** Converge
-

without the pseudo-prior, using discrete $p(x_{kt})$ may lead to convergence problem. Therefore, we approximate $p(x_{kt})$ by Gaussian $\mathcal{CN}(x_{kt}|0, \sigma_x^2)$. As a result, the approximated mean and variance are

$$\tau_{\tilde{x}_{kt}} = (\sigma_x^{-2} + \sum_l \tau_{\tilde{\mathbf{x}}_{lk}}^{-1})^{-1}; \quad m_{\tilde{x}_{kt}} = \tau_{\tilde{x}_{kt}} \sum_l \tau_{\tilde{\mathbf{x}}_{lk}}^{-1} m_{\tilde{x}_{lk}} \quad (10.27)$$

To estimate the posterior $m_{\tilde{x}_{kt}}$, We denote the posterior mean, variance and second-order moment as $m_{\tilde{x}_{kt}}, \tau_{\tilde{x}_{kt}}$ and $r_{\tilde{x}_{kt}}$.

10.4.2 Maximization Step

Investigate the log pdf inside the expectation in (10.22). By neglecting terms that do not contain $\mathbf{h}_{lk}$

$$\begin{aligned} & -\ln p(\mathbf{Y}, \tilde{\mathbf{y}}_{p,lg}, \mathbf{h}_k, \mathbf{x}_k) \\ &= \sum_l [(\mathbf{x}_k \otimes \mathbf{I}_N) \mathbf{h}_{lk} - \mathbf{y}_l]^H \mathbf{C}_{\tilde{\mathbf{v}}_{lk}}^{-1} [(\mathbf{x}_k \otimes \mathbf{I}_N) \mathbf{h}_{lk} - \mathbf{y}_l] \\ &+ (\mathbf{h}_{lk} - \tilde{\mathbf{y}}_{p,lg} + \mathbf{m}_{\tilde{\mathbf{v}}_{p,lk}})^H \mathbf{C}_{\tilde{\mathbf{v}}_{p,lk}}^{-1} (\mathbf{h}_{lk} - \tilde{\mathbf{y}}_{p,lg} + \mathbf{m}_{\tilde{\mathbf{v}}_{p,lk}}) \\ &+ \mathbf{h}_{lk}^H \boldsymbol{\Xi}_{\mathbf{h}_{lk}}^{-1} \mathbf{h}_{lk} + \text{const.} \end{aligned} \quad (10.28)$$

Therefore, $\arg \min_{\mathbf{h}_{lk}} \text{em}(\mathbf{h}_{lk} | \tilde{\mathbf{H}})$ leads to the updates

$$\begin{aligned} \mathbf{C}_{\tilde{\mathbf{h}}_{lk}} &= \left(\mathbf{C}_{\tilde{\mathbf{v}}_{p,lk}}^{-1} + \mathbf{C}_{\tilde{\mathbf{v}}_{lk0}}^{-1} \sum_t r_{\tilde{x}_{kt}} + \boldsymbol{\Xi}_{\mathbf{h}_{lk}}^{-1} \right)^{-1} \\ \mathbf{m}_{\tilde{\mathbf{h}}_{lk}} &= \mathbf{C}_{\tilde{\mathbf{h}}_{lk}} \left[\sum_t m_{\tilde{x}_{kt}}^* \mathbf{C}_{\tilde{\mathbf{v}}_{lk0}}^{-1} \mathbf{y}_{lt} + \mathbf{C}_{\tilde{\mathbf{v}}_{p,lk}}^{-1} (\tilde{\mathbf{y}}_{p,lg} - \mathbf{m}_{\tilde{\mathbf{v}}_{p,lk}}) \right] \end{aligned} \quad (10.29)$$

The above procedure is illustrated in Algorithm 10.2. To make the algorithm robust, we change the order of maximization step and expectation step.

10.5 Bilinear VB-EP

Inspired by [63], we then integrate the pseudo prior from VL-EP to the prior distribution of $\mathbf{h}_{lk}$ and $\mathbf{x}_k$ to form a enhanced prior

$$\begin{aligned} \check{p}(\mathbf{h}_{lk}) &= \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{\tilde{\mathbf{h}}_{lk}}, \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}) \\ &:= \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{\tilde{\mathbf{h}}_{lk}}, \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}) \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{0}, \boldsymbol{\Xi}_{\mathbf{h}_{lk}}) \\ \check{p}(x_{kt}) &= \check{p}(x_{kt}) p(x_{kt}) \end{aligned} \quad (10.30)$$

The additional potentials do not contribute additional information, they only serve as regularization to prevent VB-EP from drifting away from optimal solution. Furthermore, we can define the pilot-based enhanced prior as $\check{p}(\mathbf{h}_{lG_g}) \propto p(\tilde{\mathbf{y}}_{p,lg} | \mathbf{h}_{lG_g}) \check{p}(\mathbf{h}_{lG_g})$. If we introduce $\forall k, \mathbf{Z}_{lk} = \mathbf{h}_{lk} \mathbf{x}_k^T$ and $\mathbf{z}_{lk} = \text{vec}(\mathbf{Z}_{lk})$, we can verify that our current system model has exactly the same form of [61]. The VB-EP algorithm aims to find local approximation of the marginal posterior (also called belief, denoted by $b(\cdot)$) with messages $\mu_{\cdot, \cdot}(\cdot)$ of their neighboring variables. From [61], these approximations can be written as

$$\begin{aligned}
b_{f_{\mathbf{z}_l}}(\mathbf{z}_{l\{k\}}) &= p(\mathbf{Y}_l | \mathbf{z}_{l\{k\}}) \prod_k \mu_{\mathbf{z}_{lk}; f_{\mathbf{z}_l}}(\mathbf{z}_{lk}) \\
b_{f_{\mathbf{h}_{lG_g}}}(\mathbf{h}_{lG_g}) &= \check{p}(\mathbf{h}_{lG_g}) \prod_{k \in G_g} \mu_{\mathbf{h}_{lk}; f_{\mathbf{h}_{lG_g}}}(\mathbf{h}_{lk}) \\
b_{f_{\mathbf{x}_k}}(\mathbf{x}_k) &= \check{p}(\mathbf{x}_k) \mu_{\mathbf{x}_k; f_{\mathbf{x}_k}}(\mathbf{x}_k) \\
b_{\delta_{\mathbf{h}, lk}}(\mathbf{h}_{lk}) &= \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}) e^{\int \delta_{\mathbf{x}, lk}(\mathbf{x}_k) \ln \mu_{\mathbf{z}_{lk}; \delta_{lk}}(\text{vec}(\mathbf{h}_{lk} \mathbf{x}_k^T)) d\mathbf{x}_k} \\
b_{\delta_{\mathbf{x}, lk}}(\mathbf{x}_k) &= \mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k) e^{\int \delta_{\mathbf{h}, lk}(\mathbf{h}_{lk}) \ln \mu_{\mathbf{z}_{lk}; \delta_{lk}}(\text{vec}(\mathbf{h}_{lk} \mathbf{x}_k^T)) d\mathbf{h}_{lk}} \\
b_{\mathbf{z}_{lk}}(\mathbf{z}_{lk}) &= \mu_{\mathbf{z}_{lk}; f_{\mathbf{z}_l}}(\mathbf{z}_{lk}) \mu_{\mathbf{z}_{lk}; \delta_{lk}}(\mathbf{z}_{lk}) \\
b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk}) &= \mu_{\mathbf{h}_{lk}; f_{\mathbf{h}_{lG_g}}}(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}) \\
b_{\mathbf{x}_k}(\mathbf{x}_k) &= [\mu_{\mathbf{x}_k; f_{\mathbf{x}_k}}(\mathbf{x}_k) \prod_l \mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k)]^{1/L},
\end{aligned} \tag{10.31}$$

where we use $f_{\mathbf{z}_l}$, $f_{\mathbf{x}_k}$, $f_{\mathbf{h}_{lG_g}}$ to denote the likelihood of $p(\mathbf{y}_l | \mathbf{z}_{l\{k\}})$, the enhanced prior $\check{p}(\mathbf{x}_k)$ and the pilot based enhanced prior $\check{p}(\mathbf{h}_{lG_g})$. Due to the definition of $\mathbf{Z}_{lk}$, the likelihood of $\mathbf{Z}_{lk}$ given $\mathbf{h}_{lk} \mathbf{x}_k^T$ is a Dirac delta function. Therefore, in the equations above, we use δ_{lk} to denote $p(\mathbf{Z}_{lk} | \mathbf{h}_{lk} \mathbf{x}_k^T)$. The messages are the solution system of equations consisting of the following moments constraints

$$\mathbb{E}_{b_{f_{\mathbf{z}_l}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] = \mathbb{E}_{b_{\mathbf{z}_{lk}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] \tag{10.32}$$

$$\mathbb{E}_{\delta_{lk}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] = \mathbb{E}_{b_{\mathbf{z}_{lk}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] \tag{10.33}$$

$$\mathbb{E}_{b_{f_{\mathbf{h}_{lG_g}}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] = \mathbb{E}_{b_{\mathbf{h}_{lk}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] \tag{10.34}$$

$$\mathbb{E}_{b_{\delta_{lk}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] = \mathbb{E}_{b_{\mathbf{h}_{lk}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] \tag{10.35}$$

and the following marginal constraints

$$\int b_{\delta_{lk}}(\mathbf{z}_{lk}, \mathbf{h}_{lk}, \mathbf{x}_k) d\mathbf{z}_{lk} d\mathbf{h}_{lk} = b_{\mathbf{x}_k}(\mathbf{x}_k) \tag{10.36}$$

$$b_{f_{\mathbf{x}_k}}(\mathbf{x}_k) = b_{\mathbf{x}_k}(\mathbf{x}_k). \tag{10.37}$$

$$b_{\delta_{lk}}(\mathbf{z}_{lk}, \mathbf{h}_{lk}, \mathbf{x}_k) = b_{\delta_{\mathbf{h}, lk}}(\mathbf{h}_{lk}) b_{\delta_{\mathbf{x}, lk}}(\mathbf{x}_k) \delta(\mathbf{Z}_{lk} - \mathbf{h}_{lk} \mathbf{x}_k^T). \tag{10.38}$$

where $\phi(\cdot)$ denotes the first and second order moments. For all the factors f and variables θ , we say they are neighboring to each other if θ is an argument of f . The messages describing the beliefs in (10.31) are all of the form $\mu_{\theta, f}$ which can be interpreted as messages from variable to factor. However, to make the derivation easier, we define factor to variable messages $\mu_{f, \theta}$ such that

$$\forall f \in N(\theta), \mu_{\theta, f}(\theta) = \prod_{f' \in N(\theta)/\{f\}} \mu_{f', \theta}(\theta), \tag{10.39}$$

where $N(\theta)$ denotes the neighborhood around θ . As we can see, group of factor-to-variable messages is isomorphic to the group of variable-to-factor messages. Therefore, in the following context, we only describe the iterative update of factor-to-variable messages which can also be understood as the feedback messages from the factors.

10.6 Iterative Solution to VB-EP

By considering (10.37), we get $\mu_{f_{\mathbf{x}_k}} = \check{p}(\mathbf{x}_k)$. Following a similar approach, we can get the following non-trivial updates.

10.6.1 Update of $\mu_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}}$

By making (10.32) consistent while keeping all the message except for $\mu_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}} = \mu_{\mathbf{z}_{lk}; \delta_{\mathbf{z}_l}}$ fixed, we can obtain the update for $\mu_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}}$.

$$\mu_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}}(\mathbf{z}_{lk}) = \mathcal{CN}(\mathbf{z}_{lk} | \mathbf{y}_l - \sum_{k' \neq k} \mathbf{m}_{\mathbf{z}_{lk}'; f_{\mathbf{z}_l}}, \mathbf{C}_v + \sum_{k' \neq k} \mathbf{C}_{\mathbf{z}_{lk}'; f_{\mathbf{z}_l}}). \quad (10.40)$$

10.6.2 Update of $\mu_{f_{\mathbf{H}_{lg}}; \mathbf{h}_{lk}}$

The consistency of (10.34) result to the update of $\mu_{f_{\mathbf{H}_{lg}}; \mathbf{h}_{lk}} = \mu_{\mathbf{h}_{lk}; \delta_{\mathbf{h}_{lk}}}$ by considering all the other messages fixed.

$$\begin{aligned} b_{f_{\mathbf{H}_{lg}}}(\mathbf{H}_{lg}) &= p(\tilde{\mathbf{y}}_{p,lg} | \sum_{k \in G_g} \mathbf{h}_{lk}, \mathbf{C}_{\tilde{v}}) \\ &\cdot \prod_{k \in G_g} \check{p}(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk}; f_{\mathbf{H}_{lg}}}(\mathbf{h}_{lk}); \text{ and} \\ \mu_{f_{\mathbf{H}_{lg}}; \mathbf{h}_{lk}}(\mathbf{h}_{lk}) &= \frac{\int b_{f_{\mathbf{H}_{lg}}}(\mathbf{h}_{lg}) d\mathbf{h}_{l\{k' \neq k\}}}{\mu_{\mathbf{h}_{lk}; f_{\mathbf{H}_{lg}}}(\mathbf{h}_{lk})} \end{aligned} \quad (10.41)$$

Based on Gaussian reproduction lemma, we have

$$\mu_{f_{\mathbf{H}_{lg}}; \mathbf{h}_{lk}}(\mathbf{h}_{lk}) = \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{\mathbf{Y}_p | \mathbf{h}_{lk}}, \mathbf{C}_{\mathbf{Y}_p | \mathbf{h}_{lk}}) \check{p}(\mathbf{h}_{lk}) \quad (10.42)$$

where the approximated likelihood is captured by

$$\begin{aligned} \mathbf{C}_{\mathbf{Y}_p | \mathbf{h}_{lk}} &= \mathbf{C}_{\tilde{v}} + \sum_{k' \in G_g / \{k\}} \mathbf{C}_{\tilde{\mathbf{h}}_{lk'}} \\ \mathbf{m}_{\mathbf{Y}_p | \mathbf{h}_{lk}} &= \tilde{\mathbf{y}}_{p,lg} - \sum_{k' \in G_g / \{k\}} \mathbf{m}_{\tilde{\mathbf{h}}_{lk'}} \\ \mathbf{C}_{\tilde{\mathbf{h}}_{lk}} &= \left(\mathbf{C}_{\mathbf{h}_{lk}; f_{\mathbf{H}_{lg}}}^{-1} + \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}^{-1} \right)^{-1} \\ \mathbf{m}_{\tilde{\mathbf{h}}_{lk}} &= \mathbf{C}_{\tilde{\mathbf{h}}_{lk}} \left(\mathbf{C}_{\mathbf{h}_{lk}; f_{\mathbf{H}_{lg}}}^{-1} \mathbf{m}_{\mathbf{h}_{lk}; f_{\mathbf{H}_{lg}}} + \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}^{-1} \mathbf{m}_{\check{\mathbf{h}}_{lk}} \right) \end{aligned} \quad (10.43)$$

The message mean and covariance is obtained as

$$\begin{aligned} \mathbf{C}_{f_{\mathbf{H}_{lg}}; \mathbf{h}_{lk}} &= \left(\mathbf{C}_{\mathbf{Y}_p | \mathbf{h}_{lk}}^{-1} + \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}^{-1} \right)^{-1} \\ \mathbf{m}_{f_{\mathbf{H}_{lg}}; \mathbf{h}_{lk}} &= \mathbf{C}_{f_{\mathbf{H}_{lg}}; \mathbf{h}_{lk}} \left(\mathbf{C}_{\mathbf{Y}_p | \mathbf{h}_{lk}}^{-1} \mathbf{m}_{\mathbf{Y}_p | \mathbf{h}_{lk}} + \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}^{-1} \mathbf{m}_{\check{\mathbf{h}}_{lk}} \right) \end{aligned} \quad (10.44)$$

10.6.3 Update of $\mu_{\delta_{lk}; \mathbf{x}_k}$

According to (10.36), the feedback message is

$$\mu_{\delta_{lk}; \mathbf{x}_k}(\mathbf{x}_k) \propto \prod_t \mathcal{CN}(x_{kt} | \hat{m}_{\delta_{lk}; \mathbf{x}_{kt}}, \hat{\tau}_{\delta_{lk}; \mathbf{x}_{kt}}), \quad (10.45)$$

with

$$\begin{aligned} \tau_{\delta_{lk}; \mathbf{x}_{kt}} &= \text{tr} \left[\{ \mathbf{C}_{\mathbf{z}_{lk}; \delta_{lk}} \}_{tt, N}^{-1} \mathbf{R}_{b_{\delta_{lk}; \mathbf{x}_{kt}}} \right]^{-1} \\ m_{\delta_{lk}; \mathbf{x}_{kt}} &= \hat{\tau}_{\delta_{lk}; \mathbf{x}_{kt}} \mathbf{m}_{b_{\delta_{lk}; \mathbf{x}_{kt}}}^H \{ \mathbf{C}_{\mathbf{z}_{lk}; \delta_{lk}} \}_{tt, N}^{-1} \{ \mathbf{m}_{\mathbf{z}_{lk}; \delta_{lk}} \}_{t, N}, \end{aligned} \quad (10.46)$$

where $\mathbf{m}_{b_{\delta_{lk}; \mathbf{x}_{kt}}}$ and $\mathbf{R}_{b_{\delta_{lk}; \mathbf{x}_{kt}}} = \mathbf{C}_{b_{\delta_{lk}; \mathbf{x}_{kt}}} + \mathbf{m}_{b_{\delta_{lk}; \mathbf{x}_{kt}}} \mathbf{m}_{b_{\delta_{lk}; \mathbf{x}_{kt}}}^H$ denote the mean and correlation matrix of the Gaussian pdf $b_{\delta_{lk}; \mathbf{x}_{kt}}$ and the notation $\{ \cdot \}_{t_1 t_2, N}$ denotes the block entry of size $N \times N$ located in the t_1 -th block row and t_2 -th block column. From this point, we can also update the belief by

$$b_{\delta_{lk}; \mathbf{x}_k}(\mathbf{x}_k) = \mu_{\delta_{lk}; \mathbf{x}_k}(\mathbf{x}_k) \mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k). \quad (10.47)$$

10.6.4 Update of $\mu_{\delta_{lk}; \mathbf{h}_{lk}}$

Satisfying the consistency constraint (10.35) by treating $\mu_{\delta_{lk}; \mathbf{h}_{lk}} = \mu_{\mathbf{h}_{lk}; f_{\mathbf{H}_{lG_g}}}$ as the only variable and we can obtain the update equation for it. The belief $b_{\delta_{lk}; \mathbf{h}_{lk}}$ in (10.31) can be verified as Gaussian since the exponential term is quadratic. Therefore, the outbound message is computed by $\mu_{\delta_{lk}; \mathbf{h}_{lk}}(\mathbf{h}_{lk}) = b_{\delta_{lk}; \mathbf{h}_{lk}}(\mathbf{h}_{lk}) / \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk})$. The mean and covariance matrix of the Gaussian message $\mu_{\delta_{lk}; \mathbf{h}_{lk}}$ are

$$\begin{aligned} \mathbf{C}_{\delta_{lk}; \mathbf{h}_{lk}} &= \left(\sum_t [\mathbf{r}_{b_{\delta_{lk}; \mathbf{x}_{kt}}}]_t \{ \mathbf{C}_{\mathbf{z}_{lk}; \delta_{lk}} \}_{tt, N}^{-1} \right)^{-1} \\ \mathbf{m}_{\delta_{lk}; \mathbf{h}_{lk}} &= \mathbf{C}_{\delta_{lk}; \mathbf{h}_{lk}} \left(\sum_t [\mathbf{m}_{b_{\delta_{lk}; \mathbf{x}_{kt}}}]_t^* \{ \mathbf{C}_{\mathbf{z}_{lk}; \delta_{lk}} \}_{tt, N} \{ \mathbf{m}_{\mathbf{z}_{lk}; \delta_{lk}} \}_{t, N} \right), \end{aligned} \quad (10.48)$$

where $\mathbf{r}_{b_{\delta_{lk}; \mathbf{x}_{kt}}} = \mathbb{E}_{b_{\delta_{lk}; \mathbf{x}_{kt}}}[\mathbf{x}_k \cdot \mathbf{x}_k^*]$, $\mathbf{m}_{b_{\delta_{lk}; \mathbf{x}_{kt}}} = \mathbb{E}_{b_{\delta_{lk}; \mathbf{x}_{kt}}}[\mathbf{x}_k]$ with " $\cdot$ " denoting element-wise product. We update the belief by

$$\begin{aligned} b_{\delta_{lk}; \mathbf{h}_{lk}}(\mathbf{h}_{lk}) &= \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{b_{\delta_{lk}; \mathbf{h}_{lk}}}, \mathbf{C}_{b_{\delta_{lk}; \mathbf{h}_{lk}}}) \\ &= \mu_{\delta_{lk}; \mathbf{h}_{lk}}(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}). \end{aligned} \quad (10.49)$$

10.6.5 Update of $\mu_{\delta_{lk}; \mathbf{z}_{lk}}$

We examine the moments constraint given by (10.33). Based on the additional constraint (10.38), the update equation of $\mu_{\delta_{lk}; \mathbf{z}_{lk}}$ can be derived as

$$\mu_{\delta_{lk}; \mathbf{z}_{lk}}(\mathbf{z}_{lk}) = \frac{\text{proj}[b_{\delta_{lk}}(\mathbf{z}_{lk})]}{\mu_{\mathbf{z}_{lk}; \delta_{lk}}(\mathbf{z}_{lk})}, \quad (10.50)$$

where the operation $q(\mathbf{z}_{lk}) = \text{proj}[p(\mathbf{z}_{lk})]$ projects the distribution p to Gaussian family q of block covariance matrices, such that the sufficient statistics $\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})$ of p and q are the same. Thus, the message $\mu_{\delta_{lk}; \mathbf{z}_{lk}}$ is updated by

$$\begin{aligned} \mu_{\delta_{lk}; \mathbf{z}_{lk}}(\mathbf{z}_{lk}) &= \mathcal{CN}(\mathbf{z}_{lk} | \mathbf{m}_{\delta_{lk}; \mathbf{z}_{lk}}, \mathbf{C}_{\delta_{lk}; \mathbf{z}_{lk}}) \\ &= \frac{\mathcal{CN}(\mathbf{z}_{lk} | \mathbf{m}_{b_{\delta_{lk}; \mathbf{z}_{lk}}}, \mathbf{C}_{b_{\delta_{lk}; \mathbf{z}_{lk}}})}{\mathcal{CN}(\mathbf{z}_{lk} | \mathbf{m}_{\mathbf{z}_{lk}; \delta_{lk}}, \mathbf{C}_{\mathbf{z}_{lk}; \delta_{lk}})}, \end{aligned} \quad (10.51)$$

Algorithm 10.3: VB-EP [61]

- 1: Initialize: all factor-to-variable messages
 - 2: **repeat**
 - 3: $\mathbf{C}_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}}, \mathbf{m}_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}}$ by (10.40)
 - 4: $\mathbf{C}_{f_{H_{lG_g}}; \mathbf{h}_{lk}}, \mathbf{m}_{f_{H_{lG_g}}; \mathbf{h}_{lk}}$ by (10.43)-(10.44)
 - 5: $b_{\delta_{\mathbf{h}, lk}}(\mathbf{h}_{lk}) = \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}) \mu_{\delta_{lk}; \mathbf{h}_{lk}}(\mathbf{h}_{lk})$
 - 6: $\tau_{\delta_{lk}; \mathbf{x}_{kt}}, \mathbf{m}_{\delta_{lk}; \mathbf{x}_{kt}}$ by (10.46)
 - 7: $\mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k) = \tilde{p}(\mathbf{x}_k) \prod_{l' \neq l} \mu_{\delta_{l'k}; \mathbf{x}_k}(\mathbf{x}_k)$
 - 8: $b_{\delta_{\mathbf{x}, lk}}(\mathbf{x}_k) = \mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k) \mu_{\delta_{lk}; \mathbf{x}_k}(\mathbf{x}_k)$
 - 9: $\mathbf{C}_{\delta_{lk}; \mathbf{h}_{lk}}, \mathbf{m}_{\delta_{lk}; \mathbf{h}_{lk}}$ by (10.48)
 - 10: $b_{\delta_{\mathbf{h}, lk}}(\mathbf{h}_{lk}) = \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}) \mu_{\delta_{lk}; \mathbf{h}_{lk}}(\mathbf{h}_{lk})$
 - 11: $\mathbf{C}_{\delta_{lk}; \mathbf{z}_{lk}}, \mathbf{m}_{\delta_{lk}; \mathbf{z}_{lk}}$ by (10.51)
 - 12: **until** Converge
-

Algorithm 10.4: Proposed Method

- 1: [Initialize $\mathbf{U}$] $\Rightarrow$ Algorithm 10.1
 - 2: $(K, G_g, \Xi_{\mathbf{h}_{lk}}) = \text{prune}(\bar{K}, \bar{G}_g, \Xi_{\bar{\mathbf{h}}_{l\bar{k}}})$
 - 3: [Pseudo Prior] $\Rightarrow$ Algorithm 10.2
 - 4: [VB-EP] $\Rightarrow$ Algorithm 10.3
-

with

$$\begin{aligned} \mathbf{m}_{b_{\delta_{\mathbf{z}, lk}}} &= \mathbf{m}_{b_{\delta_{\mathbf{x}, lk}}} \otimes \mathbf{m}_{b_{\delta_{\mathbf{h}, lk}}} \\ \mathbf{C}_{b_{\delta_{\mathbf{z}, lk}}} &= \text{diag}(\mathbf{r}_{b_{\delta_{\mathbf{x}, lk}}}) \otimes \mathbf{C}_{b_{\delta_{\mathbf{h}, lk}}} + \mathbf{C}_{b_{\delta_{\mathbf{x}, lk}}} \otimes \mathbf{m}_{b_{\delta_{\mathbf{h}, lk}}} \mathbf{m}_{b_{\delta_{\mathbf{h}, lk}}}^H, \end{aligned}$$

where $\mathbf{C}_{b_{\delta_{\mathbf{x}, lk}}} = \mathbb{E}_{b_{\delta_{\mathbf{x}, lk}}}[(\mathbf{x}_k - \mathbf{m}_{b_{\delta_{\mathbf{x}, lk}}})(\mathbf{x}_k - \mathbf{m}_{b_{\delta_{\mathbf{x}, lk}}})^H]$ is the covariance of the belief $b_{\delta_{\mathbf{x}, lk}}$. The above procedure is concluded in **Algorithm 10.3**. Finally, by combining all the previous section together, we sum up the whole algorithm in **Algorithm 10.4**.

10.7 Simulation Results

We consider a CF MaMIMO system covering a $400 \times 400m^2$ area. A total of 16 APs are located at coordinates $\{\frac{400i}{3} + \frac{400j}{3} | i, j \in [0, 3]\}$ with a height of $10m$. The $\bar{K} = 16$ UTs are uniformly distributed in the area. To evaluate the performance of the algorithm, we generate 50 UT location setups. Following the settings in [2, 62], the data symbol power is set to $\sigma_x^2 = 14\text{dbm}$, the noise power to $\sigma_v^2 = -96\text{dbm}$ the u_k are generated by uniform distribution. The channel fading model is defined based on the distance d_{lk} between AP l and UT $\bar{k}$ as $\sigma_{h_{lk}}^2 = -30.5 - 36.7 \log_{10}(\frac{d_{lk}}{1m})$ dB. We set $P = 8$ with orthogonal pilot sequence randomly assigned to two users. The data sequence contains 4QAM symbols and has length $T = 20$. For each UT position setup, we generate 50 iterations. Consider three performance metrics, Channel Normalized Mean Squared Error(CNMSE), Activity Detection Error Rate (DER) and Symbol Error Rate (SER), For each metric, we compute the empirical cumulative distribution function (CDF) across the different position setups. We compare the proposed PP-VB-EP with the "Joint Activity Detection, Channel Estimation, and Data Detection based on expectation propagation" (JACD-EP) presented in [62]. The results of our proposed method is plotted with solid curved while the ones

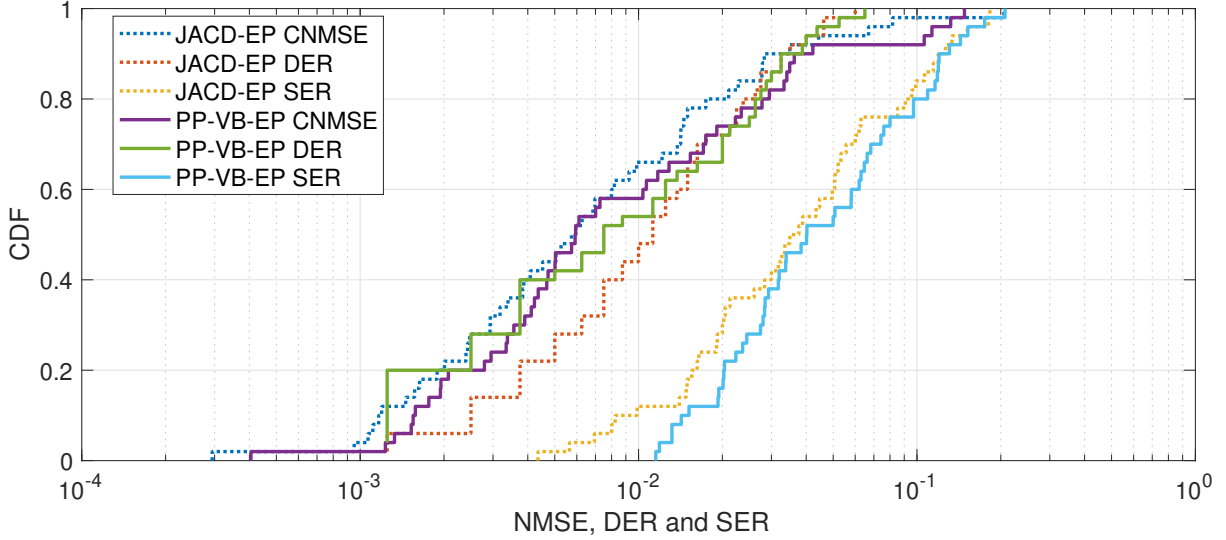


Figure 10.1: Comparison between PP-VB-EP and JACD-EP

from JACD-EP are plotted with dotted curves. The Algorithm 1 is able to give a better result compared to previous work, which verifies the correctness of our approximated log-likelihood (10.17). However, since we neglect the estimation error of $\mathbf{U}$, the estimation of $\mathbf{H}$ and $\mathbf{X}$ performs slightly worse than the previous algorithm.

10.8 Conclusions

We propose an iterative distributed Maximum Likelihood (ML) algorithm for activity detection, which simplifies the joint probability density function $p(\mathbf{Y}, \mathbf{Y}_p)$ by leveraging the CLT and large-system approximations. Building on the concept of pseudo priors, we introduce PP-VB-EP, which incorporates a distributed version of the VL-EP algorithm. Simulation results demonstrate that our proposed classical iterative distributed ML algorithm outperforms the EP-based approach in activity detection. However, due to the neglecting of detection errors of user activity, some performance degradation is observed in channel estimation and data detection. Nevertheless, VB-EP allows us to handle continuous data, and notably, its computational complexity does not scale with the constellation size.

Chapter 11

Approximating Univariate Factored Distributions via Message-Passing Algorithms

11.1 Introduction

The Gaussian Mixture Model (GMM) commonly appears in communication problems. In bilinear joint channel estimation and data detection problems, the channel likelihood can be characterized as a GMM [2]. If the data sequence length is greater than one, this GMM becomes the product of multiple simple GMM factors. As the data length increases, the number of simple GMM factors composing the product GMM grows linearly. As a result, the number of components in the product GMM grows exponentially with the number of factors.

Although the moments of a single simple GMM with a limited number of components are easy to obtain, the moments of the product GMM in the bilinear estimation and detection problems become intractable as the number of factors increases. In [23, 56, 64], the authors attempt to avoid GMMs in bilinear joint estimation and detection using various methods. Distributions in the form of a product of GMMs can be generalized as univariate factored distributions. This paper investigates iterative message-passing-based algorithms to approximate the statistics of univariate factored distributions.

11.1.1 Prior Works

Given a univariate factored distribution, message-passing algorithms can be used to obtain the marginal distributions, which is equivalent to finding a stationary point of a constrained Bethe Free Energy (BFE) minimization [11].

Belief Propagation

Belief Propagation (BP) was introduced by Pearl as an approximate inference tool [65]. Within the BFE framework, BP aims to find a stationary point of the BFE under strict marginal consistency constraints. BP is defined on the basis of a factored probability density function (PDF). A corresponding factor graph can be constructed by representing all factors and variables as nodes, and connecting a factor node to a variable node whenever the variable appears as an argument of that factor. By iteratively applying the

sum-product rules, BP approximates the marginal distributions via auxiliary functions, called messages [5, 66].

From the constrained BFE optimization perspective, BP seeks a stationary point of the BFE under strict marginal consistency constraints. This requirement often results in intractable functional forms for the messages. Consequently, even though BP reduces the dimensionality of the inference problem, its computational complexity may still be prohibitive when the messages are intractable—for example, when the factored pdf is a product of GMMs. [67, 68]

Gaussian BP (GaBP) [64, 69] is a variant of BP that projects messages to Gaussian distributions based on the Central Limit Theorem (CLT).

Expectation Propagation

Expectation Propagation (EP) [7] was proposed to handle intractable messages by introducing a projection step that approximates messages within a chosen family of distributions. In the constrained BFE minimization framework, this corresponds to relaxing the strict marginal consistency constraints to moment constraints [11].

Compared to BP, a major problem with EP is that the factor-level beliefs (i.e., approximated marginals involving the arguments of a factor) may be non-integrable. This prevents EP from proceeding, since the computation of means and variances of such beliefs is required. In [70], the authors investigated this problem and proposed several methods to address it.

11.1.2 Main Contribution

This paper investigates both BP- and EP-based approaches for estimating the mean and variance of a univariate factored distribution.

BP cannot be directly applied to approximate these statistics when every factor is, for example, a GMM. To address this issue, we propose Variable Duplication Gaussian Belief Propagation (VDBP) to avoid intractable computations. In multivariate inference problems, BP provides marginal distributions for individual variables, as well as joint marginals for variables appearing together in a factor. Leveraging this property, we construct a multivariate linear measurement model whose underlying marginal posterior (or joint PDF with all but one variable marginalized) is equivalent to the given univariate factored distribution up to a constant factor. We then apply GaBP to iteratively approximate the marginal posterior of the multivariate linear measurement model.

As the name suggests, VDBP converts a univariate factored PDF into a multivariate factored PDF by duplicating one variable for each factor, treating the univariate factored distribution as a multivariate joint PDF in which each factor is proportional to the PDF of an independent random variable. The resulting multivariate factored PDF is interpreted as prior information. The measurement channel and measurement outcome are designed such that the measurement likelihood enforces equality among all duplicated variables. Based on the properties of Dirac delta functions, VDBP constructs the measurement matrix as the transpose of the orthogonal complement of the all-ones vector (i.e., each row of the measurement matrix is orthogonal to the all-ones vector). The measurement outcome and noise are both constructed as all-zero vectors. We show analytically that the marginal posterior of the constructed measurement model is equivalent to the original univariate factored PDF.

To overcome the problem of non-integrable beliefs when EP is used to approximate the statistics of univariate factored distributions, we adopt ideas from recent work [70]. We investigated a persistent sequential EP algorithm for estimating the univariate factored PDF, in which updates associated with non-integrable beliefs are skipped.

Furthermore, inspired by [70], we propose Analytic Continuation EP (ACEP), in which Gaussian distributions are represented using natural parameters. The technique of analytic continuation allows us to exploit discontinuities in Gaussian functions expressed in natural-parameter form. The key idea of ACEP is to dynamically shrink the projection family in EP to ensure that the belief associated with the next factor is integrable.

Both EP-based algorithms proposed in this paper operate in a sequential manner. We show that the variable-level beliefs produced by both EP based algorithms are guaranteed to be integrable.

11.2 Problem Formulation

Consider a univariate factored PDF with non-negative integrable factors $f_n(\theta)$

$$p(\theta) \propto \prod_{n=1}^N f_n(\theta), \quad (11.1)$$

where the objective of this paper is to develop a low-complexity method for estimating the statistics of θ . The direct computation of integrals involving $p(\theta)$ becomes intractable when each factor f_n is a mixture model.

11.3 VDBP-Based Approximation for Univariate Factored PDFs

11.3.1 Outline of the VDBP

Inspired by [19], we duplicate the variable θ into N new variables $\{\theta_n\}_{n=1}^N$, one for each factor f_n , to construct a multivariate joint PDF. To describe the equality relationships among these newly constructed variables, we define a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ with $M = N - 1$, such that $\mathbf{A} \cdot \mathbf{1} = \mathbf{0}$. This matrix can be constructed by first generating a full-rank random matrix $\mathbf{B} \in \mathbb{R}^{M \times N}$ and then setting

$$\mathbf{A} = \mathbf{B} - \text{diag} \left(\frac{\mathbf{B} \cdot \mathbf{1}_N}{N} \right) \mathbf{1}_{N-1} \mathbf{1}_N^T. \quad (11.2)$$

A simple choice for such a matrix is a trimmed Hadamard matrix with the all-ones row removed.

We define the multivariate factored joint PDF as

$$p(\boldsymbol{\theta}) \propto \delta(\mathbf{A}\boldsymbol{\theta}) \prod_n f_n(\theta_n) \propto \prod_m \delta(\mathbf{A}_{m,:} \boldsymbol{\theta}) \prod_n f_n(\theta_n), \quad (11.3)$$

where $\mathbf{A}_{m,:}$ denotes the m -th row of the matrix $\mathbf{A}$.

Lemma 5. $\forall \mathbf{A} \in \mathbb{R}^{(N-1) \times N}$ and $\forall n \in \{1, \dots, N\}$, let $\overline{\mathbf{A}}_n$ denote the matrix obtained by removing the n -th column of $\mathbf{A}$. If $\text{rank}(\mathbf{A}) = N - 1$ and $\mathbf{A} \cdot \mathbf{1} = \mathbf{0}$, then $\text{rank}(\overline{\mathbf{A}}_n) = N - 1$.

Proof. We prove the lemma by contradiction. Assume that $\text{rank}(\overline{\mathbf{A}}_n) < N - 1$, then $\exists k \in \{1, \dots, N - 1\}$ and $\exists \mathbf{b}^T \in \mathbb{R}^{1 \times (N-2)}$, such that

$$[\overline{\mathbf{A}}_n]_{k,:} = \mathbf{b}^T \mathbf{C}_{kn}, \quad (11.4)$$

where $\mathbf{C}_{kn} \in \mathbb{R}^{(N-2) \times (N-1)}$ is defined as

$$\mathbf{C}_{kn} = \begin{bmatrix} [\overline{\mathbf{A}}_n]_{1:(k-1),:} \\ [\overline{\mathbf{A}}_n]_{(k+1):(N-1),:} \end{bmatrix}. \quad (11.5)$$

The above two equations indicate that the k -th row of $\overline{\mathbf{A}}_n$ can be written as a linear combination of the remaining rows in $\overline{\mathbf{A}}_n$.

On the other hand, due to the assumption $\mathbf{A} \cdot \mathbf{1} = \mathbf{0}$, we have

$$\overline{\mathbf{A}}_n \cdot \mathbf{1} = -\mathbf{a}_n, \quad (11.6)$$

where $\mathbf{a}_n$ denotes the n -th column of $\mathbf{A}$. We can remove the k -th rows on both sides of (11.6):

$$\mathbf{C}_{kn} \cdot \mathbf{1} = -\mathbf{d}_{kn}, \quad (11.7)$$

where $\mathbf{d}_{kn}$ is a trimmed version of $\mathbf{a}_n$, with the k -th entry removed. Similarly, by extracting the k -th row in (11.6) and we have

$$[\overline{\mathbf{A}}_n]_{k,:} \cdot \mathbf{1} = -[\mathbf{A}]_{k,n}. \quad (11.8)$$

Right-multiplying both sides of (11.4) by the all-ones vector and using (11.7) and (11.8), we obtain

$$\begin{aligned} [\overline{\mathbf{A}}_n]_{k,:} \cdot \mathbf{1} &= \mathbf{b}^T \mathbf{C}_{kn} \cdot \mathbf{1} \\ \Rightarrow [\mathbf{A}]_{k,n} &= \mathbf{b}^T \mathbf{d}_{kn}. \end{aligned} \quad (11.9)$$

Combining (11.4) and (11.9), we have

$$\begin{aligned} \begin{bmatrix} [\mathbf{A}]_{k,n} & [\overline{\mathbf{A}}_n]_{k,:} \end{bmatrix} &= \mathbf{b}^T \begin{bmatrix} \mathbf{d}_{kn} & \mathbf{C}_{kn} \end{bmatrix} \\ \Rightarrow \begin{bmatrix} [\mathbf{A}]_{k,n} & [\overline{\mathbf{A}}_n]_{k,:} \end{bmatrix} \mathbf{P} &= \mathbf{b}^T \begin{bmatrix} \mathbf{d}_{kn} & \mathbf{C}_{kn} \end{bmatrix} \mathbf{P}, \end{aligned} \quad (11.10)$$

where $\mathbf{P}$ is a $N \times N$ permutation matrix:

$$\mathbf{P} = \begin{bmatrix} \mathbf{0}_{n-1}^T & 1 & \mathbf{0}_{N-n}^T \\ \mathbf{I}_{n-1} & \mathbf{0}_{n-1} & \mathbf{O}_{(n-1) \times (N-n)} \\ \mathbf{O}_{(N-n) \times (n-1)} & \mathbf{0}_{N-n} & \mathbf{I}_{N-n} \end{bmatrix}. \quad (11.11)$$

One can verify that $\begin{bmatrix} [\mathbf{A}]_{k,n} & [\overline{\mathbf{A}}_n]_{k,:} \end{bmatrix} \mathbf{P}$ is the k -th row of $\mathbf{A}$ and that $\begin{bmatrix} \mathbf{d}_{kn} & \mathbf{C}_{kn} \end{bmatrix} \mathbf{P}$ is the matrix obtained from $\mathbf{A}$ by removing its k -th row. Therefore, (11.10) indicates $\text{rank}(\mathbf{A}) < N - 1$, which contradicts the assumption that $\text{rank}(\mathbf{A}) = N - 1$. $\square$

Theorem 2. $\forall \mathbf{A} \in \mathbb{R}^{M \times N}$, where $M = N - 1$, if $\text{rank}(\mathbf{A}) = M$ and $\mathbf{A} \cdot \mathbf{1} = \mathbf{0}$, then

$$\forall n : \int \delta(\mathbf{A}\boldsymbol{\theta}) \prod_{n' \neq n} f_{n'}(\theta_{n'}) d\theta_{n'} \propto \prod_{n'=1}^N f_{n'}(\theta_n). \quad (11.12)$$

Proof. Consider the term inside the Dirac delta function

$$\delta(\mathbf{A}\boldsymbol{\theta}) = \delta(\overline{\mathbf{A}}_n \overline{\boldsymbol{\theta}}_n + \mathbf{a}_n \theta_n), \quad (11.13)$$

where $\overline{\mathbf{A}}_n$ denotes the $M \times M$ matrix obtained from $\mathbf{A}$ by removing its n -th column. Similarly, $\overline{\boldsymbol{\theta}}_n$ is the vector obtained from $\boldsymbol{\theta}$ by removing its n -th entry. Furthermore, $\mathbf{a}_n$ denotes the n -th column of $\mathbf{A}$ and θ_n denotes the n -th entry of $\boldsymbol{\theta}$.

By Lemma 5, $\overline{\mathbf{A}}_n$ is an invertible matrix. Define $\boldsymbol{\eta} = \overline{\mathbf{A}}_n \overline{\boldsymbol{\theta}}_n$, and thus, $d\overline{\boldsymbol{\theta}}_n = \left| \det \left(\overline{\mathbf{A}}_n^{-1} \right) \right| d\boldsymbol{\eta}$. Therefore, the integral in the theorem can be rewritten as

$$\begin{aligned} & \int \delta(\mathbf{A}\boldsymbol{\theta}) \prod_{n' \neq n} f_{n'}(\theta_{n'}) d\theta_{n'} \\ &= \frac{1}{|\det(\overline{\mathbf{A}}_n)|} \int d\boldsymbol{\eta} \delta(\boldsymbol{\eta} + \mathbf{a}_n \theta_n) \prod_{n' \neq n} f_{n'} \left(\left[\overline{\mathbf{A}}_n^{-1} \right]_{k(n'), :} \boldsymbol{\eta} \right) \\ &= \frac{1}{|\det(\overline{\mathbf{A}}_n)|} \prod_{n' \neq n} f_{n'} \left(- \left[\overline{\mathbf{A}}_n^{-1} \right]_{k(n'), :} \mathbf{a}_n \theta_n \right), \end{aligned} \quad (11.14)$$

where $k(n') : \{1, \dots, n-1, n+1, \dots, N\} \rightarrow \{1, \dots, N-1\}$ is an index transformation defined as

$$k(n') = \begin{cases} n' & \text{if } 1 \leq n' < n \\ n' - 1 & \text{if } n < n' \leq N \end{cases}. \quad (11.15)$$

On the other hand, from the assumption $\mathbf{A} \cdot \mathbf{1} = \mathbf{0}$, we have

$$\begin{aligned} & \mathbf{A} \cdot \mathbf{1} = \mathbf{0} \\ & \Rightarrow \overline{\mathbf{A}}_n \cdot \mathbf{1} = -\mathbf{a}_n \\ & \Rightarrow \mathbf{1} = -\overline{\mathbf{A}}_n^{-1} \mathbf{a}_n \\ & \Rightarrow \forall j \in \{1, \dots, N-1\} : \left[\overline{\mathbf{A}}_n^{-1} \right]_{j, :} \mathbf{a}_n = -1. \end{aligned} \quad (11.16)$$

Substitute (11.16) into the last line of (11.14) yields

$$\int \delta(\mathbf{A}\boldsymbol{\theta}) \prod_{n' \neq n} f_{n'}(\theta_{n'}) d\theta_{n'} = \frac{1}{|\det(\overline{\mathbf{A}}_n)|} \prod_{n'=1}^N f_{n'}(\theta_n). \quad (11.17)$$

□

Based on Theorem 2, we can view (11.1) as the marginalized version of (11.3), i.e.,

$$\forall i \in \mathbb{N} \cap [1, N] : \prod_n f_n(\theta_i) \propto \int \delta(\mathbf{A}\boldsymbol{\theta}) \prod_n f_n(\theta_n) d\overline{\boldsymbol{\theta}}_i, \quad (11.18)$$

where $\overline{\boldsymbol{\theta}}_i$ denotes an $(N-1) \times 1$ vector obtained by removing the i -th entry from $\boldsymbol{\theta}$. The delta distribution $\delta(\mathbf{A}\boldsymbol{\theta})$ can be interpreted as a noiseless linear measurement model:

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{v}, \quad (11.19)$$

where $\mathbf{y} = \mathbf{0}$ and $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \text{diag}([\epsilon_1 \dots \epsilon_N]))$, with infinitesimal ϵ_n for all n . The auxiliary noise variance ϵ_n can serve as a regularization term to prevent zero denominators

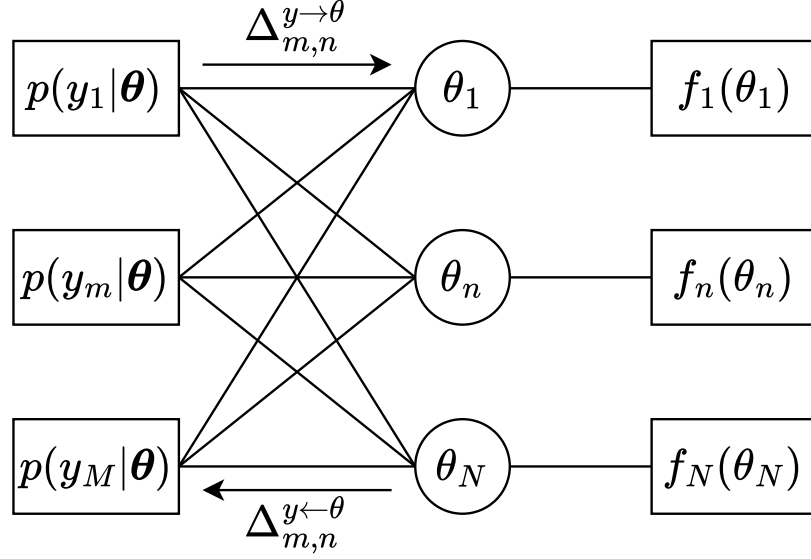


Figure 11.1: Factor Graph for VDBP

that may occur due to poor initialization. In our simulations, we set $\epsilon_n = 0$. The corresponding factor graph for (11.3) and (11.19) is illustrated in Figure 11.1.

Since each factor node $f_i(\theta_i)$ is connected to only one variable θ_i , the message from factor f_i to θ_i is always $f_i(\theta_i)$. A tractable BP-based algorithm for estimating the statistics can be obtained by deriving the message update equations for $\Delta_{m,n}^{y \rightarrow \theta}$ and $\Delta_{m,n}^{y \leftarrow \theta}$. These messages can also be interpreted as approximate probabilities: the factor-to-variable message $\Delta_{m,n}^{y \rightarrow \theta}$ is proportional to the approximate likelihood $p(y_m|\theta_n)$, while the variable-to-factor message $\Delta_{m,n}^{y \leftarrow \theta}$ is proportional to the approximate posterior $p(\theta_n|\bar{\mathbf{y}}_m)$, where $\bar{\mathbf{y}}_m$ denotes all entries of $\mathbf{y}$ except the m -th one.

In the following, we apply the BP update rules, together with the CLT, to derive these messages.

11.3.2 Derivation of the VDBP

We first derive the factor-to-variable message $\Delta_{m,n}^{y \rightarrow \theta}(\theta_n)$. Since BP is an iterative method, we assume that the mean and variance of the variable-to-factor message $\Delta_{m,n}^{y \leftarrow \theta}(\theta_n)$ are given by $\mu_{m,n}^\theta = \mathbb{E}_{\Delta_{m,n}^{y \leftarrow \theta}}[\theta_n]$ and $\tau_{m,n}^\theta = \text{Var}_{\Delta_{m,n}^{y \leftarrow \theta}}[\theta_n]$, respectively.

Define $z_m = \sum_n a_{mn}\theta_n$. Then the conditional PDF $p(z_m|\theta_n, \{\Delta_{m,n}^{y \leftarrow \theta}\}_{n=1}^N)$ can be approximated as Gaussian according to the CLT

$$\begin{aligned}
& p(z_m|\theta_n, \{\Delta_{m,n}^{y \leftarrow \theta}\}_{n=1}^N) \\
&= \int \delta(z_m - a_{mn}\theta_n - \sum_{n' \neq n} a_{mn'}\theta_{n'}) \prod_{n' \neq n} \Delta_{m,n'}^{y \leftarrow \theta}(\theta_{n'}) d\theta_{n'} \\
&\simeq \mathcal{N}(z_m|a_{mn}\theta_n + \mu_{n,m}^p, \tau_{n,m}^p),
\end{aligned} \tag{11.20}$$

where $\{\Delta_{m,n}^{y \leftarrow \theta}\}_{n=1}^N$ denotes the collection of all variable-to-factor messages. In the last line of (11.20), the CLT is applied to approximate the interference term $\sum_{n' \neq n} a_{mn'}\theta_{n'}$ as

Gaussian with mean and variance $\mu_{n,m}^p, \tau_{n,m}^p$:

$$\begin{aligned}\mu_{n,m}^p &= \sum_{n' \neq n} a_{mn'} \mu_{m,n'}^\theta \\ \tau_{n,m}^p &= \sum_{n' \neq n} |a_{mn'}|^2 \tau_{m,n'}^\theta.\end{aligned}\tag{11.21}$$

Based on BP rules for the feedback messages, we combine $p(z_m|\theta_n, \{\Delta_{m,n}^{y \leftarrow \theta}\}_{n=1}^N)$ with the measurement model $p(y_m|z_m) = \mathcal{N}(y_m|z_m, \epsilon_m)$ and integrate over z_m to obtain the factor-to-variable message $\Delta_{m,n}^{y \rightarrow \theta}(\theta_n)$, which corresponds to the approximate likelihood $p(y_m|\theta_n)$

$$\Delta_{m,n}^{y \rightarrow \theta}(\theta_n) \propto \mathcal{N}\left(\theta_n \left| \frac{y_m - \mu_{n,m}^p}{a_{mn}}, \frac{\tau_{n,m}^p + \epsilon_m}{|a_{mn}|^2} \right. \right).\tag{11.22}$$

We now derive the variable-to-factor message $\Delta_{m,n}^{y \leftarrow \theta}$:

$$\begin{aligned}\Delta_{m,n}^{y \leftarrow \theta}(\theta_n) &\propto f_n(\theta_n) \prod_{j \neq m} p(y_j|\theta_n) \\ &= f_n(\theta_n) \mathcal{N}(\theta_n | \mu_{m,n}^r, \tau_{m,n}^r),\end{aligned}\tag{11.23}$$

where $\mu_{m,n}^r$ and $\tau_{m,n}^r$ follow from the Gaussian reproduction property [15]:

$$\begin{aligned}\tau_{m,n}^r &= \left(\sum_{j \neq m} \frac{|a_{jn}|^2}{\tau_{n,j}^p + \epsilon_j} \right)^{-1} \\ \mu_{m,n}^r &= \tau_{m,n}^r \left(\sum_{j \neq m} \frac{a_{jn} (y_j - \mu_{n,j}^p)}{\tau_{n,j}^p + \epsilon_j} \right).\end{aligned}\tag{11.24}$$

From (11.20)–(11.21), we observe that under the CLT approximation only the first two moments of $\Delta_{m,n}^{y \leftarrow \theta}$ are required:

$$\begin{aligned}\mu_{m,n}^\theta &= \mathbb{E}_{\Delta_{m,n}^{y \leftarrow \theta}}[\theta_n] \\ \tau_{m,n}^\theta &= \text{Var}_{\Delta_{m,n}^{y \leftarrow \theta}}[\theta_n],\end{aligned}\tag{11.25}$$

where $\Delta_{m,n}^{y \leftarrow \theta}$ is defined in (11.23). This concludes our discussion for a general factor f_n in (11.1). Next, we consider a representative case in which f_n follows a GMM that arises in bilinear joint estimation and detection problems.

GMM Factors

Assume that the factors in (11.1) have the following structure:

$$f_n(\theta_n) \propto \sum_{s_n} p_{s_n} \mathcal{N}(\theta_n | \mu_{s_n}, \tau_{s_n}).\tag{11.26}$$

The only part of VDBP that depends on the factor structure is the update of the mean and variance of $\Delta_{m,n}^{y \leftarrow \theta}$ in (11.25). Substituting (11.26) into (11.23) and applying the Gaussian Reproduction Lemma [15], we obtain

$$\begin{aligned}\Delta_{m,n}^{y \leftarrow \theta}(\theta_n) &\propto \sum_{s_n} \pi_n(s_n | \mu_{m,n}^r, \tau_{m,n}^r) \\ &\cdot \mathcal{N}(\theta_n | \mu_{m,n}^{x|s}(s_n), \tau_{m,n}^{x|s}(s_n)),\end{aligned}\tag{11.27}$$

where $\pi_n(s_n|\mu_{m,n}^r, \tau_{m,n}^r)$ is the normalized categorical distribution defined as

$$\pi_n(s_n|\mu_{m,n}^r, \tau_{m,n}^r) \propto p_{s_n} \mathcal{N}(0|\mu_{s_n} - \mu_{m,n}^r, \tau_{m,n}^r + \tau_{s_n}). \quad (11.28)$$

The corresponding conditional mean and variance are given by

$$\begin{aligned} \mu_{m,n}^{x|s}(s_n) &= \frac{\tau_{m,n}^r \mu_{s_n} + \tau_{s_n} \mu_{m,n}^r}{\tau_{m,n}^r + \tau_{s_n}} \\ \tau_{m,n}^{x|s}(s_n) &= \frac{\tau_{m,n}^r \tau_{s_n}}{\tau_{m,n}^r + \tau_{s_n}}. \end{aligned} \quad (11.29)$$

Therefore, the message $\Delta_{m,n}^{y \leftarrow \theta}(\theta_n)$ in (11.27) is itself a GMM with the same number of components as the factor f_n . By exchanging the order of integration and summation, the mean of $\Delta_{m,n}^{y \leftarrow \theta}(\theta_n)$ can be computed as

$$\mu_{m,n}^\theta = \mathbb{E}_{\pi_n(s_n|\mu_{m,n}^r, \tau_{m,n}^r)} [\mu_{m,n}^{x|s}(s_n)], \quad (11.30)$$

while its variance is given by

$$\tau_{m,n}^\theta = \mathbb{E}_{\pi_n(s_n|\mu_{m,n}^r, \tau_{m,n}^r)} [\tau_{m,n}^{x|s}(s_n) + \mu_{m,n}^{x|s}(s_n)^2] - (\mu_{m,n}^\theta)^2. \quad (11.31)$$

11.3.3 Approximated Statistics

Recall the relation in (11.18), which indicates that the univariate factorized PDF in (11.1) can be approximated by the marginal distribution of (11.3). According to the BP rule for variable level belief, the marginal distribution of a variable can be approximated as the product of all incoming messages to that variable. Since the message $\Delta_{m,n}^{y \rightarrow \theta}$ in (11.22) follows a Gaussian distribution, we define the extrinsic mean and variance μ_n^r and τ_n^r such that

$$\mathcal{N}(\theta_n|\mu_n^r, \tau_n^r) : \propto \prod_m \Delta_{m,n}^{y \rightarrow \theta}(\theta_n). \quad (11.32)$$

As a result, the extrinsic mean and variance are

$$\begin{aligned} \tau_n^r &= \left(\sum_m \frac{|a_{mn}|^2}{\tau_{n,m}^p + \epsilon_m} \right)^{-1} \\ \mu_n^r &= \tau_n^r \left(\sum_m \frac{a_{mn} (y_m - \mu_{n,m}^p)}{\tau_{n,m}^p + \epsilon_m} \right). \end{aligned} \quad (11.33)$$

By combining the extrinsic distribution in (11.32) with the factor $f_n(\theta_n)$, we obtain the approximate marginal of (11.3) as

$$b_{\theta_n}(\theta_n) : \propto f_n(\theta_n) \mathcal{N}(\theta_n|\mu_n^r, \tau_n^r). \quad (11.34)$$

According to (11.18), (11.34) can be regarded as an approximate form of (11.1). Without loss of generality, we restrict our attention to the approximate mean and variance.

Each θ_n yields its own belief of the marginal mean and variance, denoted by μ_n^θ and τ_n^θ :

$$\begin{aligned} \mu_n^\theta &= \mathbb{E}_{b_{\theta_n}}[\theta_n] \\ \tau_n^\theta &= \text{Var}_{b_{\theta_n}}[\theta_n]. \end{aligned} \quad (11.35)$$

Algorithm 11.1: VDBP for GMM Factors

- 1: Initialize $\forall m, n : \mu_{m,n}^\theta = 0, \tau_{m,n}^\theta = 1$
 - 2: **repeat**
 - 3: $\forall m : \mu_m^p = \sum_n a_{mn} \mu_{m,n}^\theta$
 - 4: $\forall m : \tau_m^p = \sum_n |a_{mn}|^2 \tau_{m,n}^\theta$
 - 5: $\forall n, m : \mu_{n,m}^p = \mu_m^p - a_{mn} \mu_{m,n}^\theta$
 - 6: $\forall n, m : \tau_{n,m}^p = \tau_m^p - |a_{mn}|^2 \tau_{m,n}^\theta$
 - 7: $\forall n : \tau_n^r = \left(\sum_m \frac{|a_{mn}|^2}{\tau_{n,m}^p + \epsilon_m} \right)^{-1}$
 - 8: $\forall n : \mu_n^r = \tau_n^r \left(\sum_m \frac{a_{mn}^* (y_m - \mu_{n,m}^p)}{\tau_{n,m}^p + \epsilon_m} \right)$
 - 9: $\forall m, n : \tau_{m,n}^r = \left(\tau_n^{r-1} - \frac{|a_{mn}|^2}{\tau_{n,m}^p + \epsilon_m} \right)^{-1}$
 - 10: $\forall m, n : \mu_{m,n}^r = \tau_{m,n}^r \left(\frac{\mu_n^r}{\tau_n^r} - \frac{a_{mn}^* (y_m - \mu_{n,m}^p)}{\tau_{n,m}^p + \epsilon_m} \right)$
 - 11: $\forall m, n : \text{Compute normalized } \pi_n(s_n) \text{ via (11.28)}$
 - 12: $\forall m, n : \text{Compute } \mu_{m,n}^\theta, \tau_{m,n}^\theta \text{ via (11.30), (11.31)}$
 - 13: **until** Convergence
-

Finally, we combine these beliefs to obtain the approximate mean and variance of $p(\theta)$ in (11.1), denoted by $\mu^{\hat{\theta}}$ and $\tau^{\hat{\theta}}$:

$$\begin{aligned} \mu^{\hat{\theta}} &= \left(\sum_n \frac{1}{\tau_n^\theta} \right)^{-1} \left(\sum_n \frac{\mu_n^\theta}{\tau_n^\theta} \right) \\ \tau^{\hat{\theta}} &= \min(\tau_1^\theta, \dots, \tau_N^\theta). \end{aligned} \tag{11.36}$$

The computational complexity of the proposed algorithm depends on the number of nonzero elements in $\mathbf{A}$. When a trimmed Hadamard matrix is used, the overall computational complexity is $O(N^2)$. The VDBP algorithm is summarized in Algorithm 11.1.

11.4 Preliminaries for Expectation Propagation

In EP, a message does not necessarily correspond to a normalizable function and may even be non-integrable. Let $\mathcal{UN}(\theta|\mu, \tau)$ denote an unnormalized Gaussian (ignoring the $\sqrt{2\pi\tau}$ normalization factor),

$$\mathcal{UN}(\theta|\mu, \tau) = \exp - \frac{(\theta - \mu)^2}{2\tau}. \tag{11.37}$$

We define a Gaussian distribution in natural-parameter form as

$$\mathcal{M}(\theta|\nu, \xi) = \frac{\sqrt{\xi}}{\sqrt{2\pi}} \exp \left(-\frac{\xi}{2} \theta^2 + \nu \theta - \frac{\nu^2}{2\xi} \right). \tag{11.38}$$

Similarly, the unnormalized Gaussian in natural-parameter form is defined as

$$\mathcal{UM}(\theta|\nu, \xi) = \exp \left(-\frac{\xi}{2} \theta^2 + \nu \theta - \frac{\nu^2}{2\xi} \right). \tag{11.39}$$

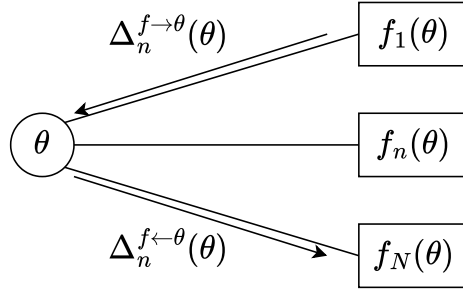


Figure 11.2: Factor Graph for EP

One can verify that if $\mathcal{UN}(\theta|\mu, \tau) = \mathcal{UM}(\theta|\nu, \xi)$, or if $\mathcal{N}(\theta|\mu, \tau) = \mathcal{M}(\theta|\nu, \xi)$, then

$$\begin{aligned}\nu &= \frac{\mu}{\tau}; \\ \xi &= \frac{1}{\tau}.\end{aligned}\tag{11.40}$$

We can then obtain the unnormalized Gaussian reproduction lemma. For all $\mu \in \mathbb{R}$ and $\tau \in \mathbb{R}/\{0\}$,

$$\begin{aligned}&\mathcal{UN}(\theta|\mu_1, \tau_1)\mathcal{UN}(\theta|\mu_2, \tau_2) \\ &= \mathcal{UN}(0|\mu_1 - \mu_2, \tau_1 + \tau_2)\mathcal{UN}\left(\theta \left| \frac{\tau_2\mu_1 + \tau_1\mu_2}{\tau_1 + \tau_2}, \frac{\tau_1\tau_2}{\tau_1 + \tau_2} \right. \right).\end{aligned}\tag{11.41}$$

Similarly, apply the Gaussian reproduction lemma to unnormalized Gaussians in natural-parameter form, and using analytic continuation to cover the case $\xi = 0$ (i.e., $\tau \rightarrow \infty$), we have $\forall \nu, \xi \in \mathbb{R}$,

$$\begin{aligned}&\mathcal{UM}(\theta|\nu_1, \xi_1)\mathcal{UM}(\theta|\nu_2, \xi_2) = \mathcal{UN}(\theta|\mu_1, \tau_1)\mathcal{UN}(\theta|\mu_2, \tau_2) \\ &= \mathcal{UM}\left(0 \left| \frac{\xi_2\nu_1 - \xi_1\nu_2}{\xi_1 + \xi_2}, \frac{\xi_1\xi_2}{\xi_1 + \xi_2} \right. \right)\mathcal{UM}(\theta|\nu_1 + \nu_2, \xi_1 + \xi_2),\end{aligned}\tag{11.42}$$

where

$$\begin{aligned}\nu_1 &= \frac{\mu_1}{\tau_1}, \nu_2 = \frac{\mu_2}{\tau_2}; \\ \xi_1 &= \frac{1}{\tau_1}, \xi_2 = \frac{1}{\tau_2}.\end{aligned}\tag{11.43}$$

11.4.1 Factor Graph of the Given Problem in EP

Now we investigate EP-based algorithms. Consider the univariate factored PDF given in (11.1). We define the message from the n -th factor to θ as $\Delta_n^{f \rightarrow \theta}(\theta)$ and the message from θ to the n -th factor as $\Delta_n^{f \leftarrow \theta}(\theta)$. The corresponding factor graph is shown in Fig. 11.2.

11.5 Persistent EP

In EP, the messages are not required to be integrable. Unlike the Gaussian distribution, the unnormalized Gaussian defined in (11.37) allows negative values of τ . However, the

point $\tau = 0$ is an essential discontinuity of (11.37). Persistent EP operates under a sequential update order. Specifically, the messages to and from one factor are updated at a time, and the latest messages are immediately used when updating the messages associated with the next factor. In the following, we focus on deriving the messages to and from factor f_n .

Persistent EP does not actively prevent the formation of non-integrable beliefs. Instead, it handles non-integrable beliefs in a passive manner: if the belief at the factor $f_n(\theta)$ is non-integrable, the update of the message $\Delta_n^{f \leftarrow \theta}(\theta)$ is skipped.

Based on the EP rules, the message from variable to factor $f_n(\theta)$ can be computed as

$$\Delta_n^{f \leftarrow \theta}(\theta) = \mathcal{UN}(\theta | \mu_n^{f \leftarrow \theta}, \tau_n^{f \leftarrow \theta}) \propto \prod_{n' \neq n} \Delta_{n'}^{f \rightarrow \theta}(\theta), \quad (11.44)$$

where

$$\begin{aligned} \tau_n^{f \leftarrow \theta} &= \left(\sum_{n' \neq n} \frac{1}{\tau_{n'}^{f \rightarrow \theta}} \right)^{-1} \\ \mu_n^{f \leftarrow \theta} &= \tau_n^{f \leftarrow \theta} \left(\sum_{n' \neq n} \frac{\mu_{n'}^{f \rightarrow \theta}}{\tau_{n'}^{f \rightarrow \theta}} \right). \end{aligned} \quad (11.45)$$

Unlike standard EP algorithm, a checking mechanism is introduced before computing the belief at factor f_n . Specifically, we evaluate whether the integral

$$\int f_n(\theta) \Delta_n^{f \leftarrow \theta}(\theta) d\theta$$

is finite. The check passes if the integral exist in $\mathbb{R}$.

If the check fails, the update of $\Delta_n^{f \rightarrow \theta}(\theta)$ is skipped in the current iteration, and the algorithm proceeds to the next factor f_k , where $k = (n \bmod N) + 1$, and repeats from (11.44) with index n replaced by k .

If the check passes, the belief at f_n can be computed as

$$b_{f_n}(\theta) \propto f_n(\theta) \Delta_n^{f \leftarrow \theta}(\theta), \quad (11.46)$$

which is guaranteed to have a finite integral value. The corresponding belief mean and variance are

$$\begin{aligned} \mu_n^{b_f} &= \mathbb{E}_{b_{f_n}}[\theta], \\ \tau_n^{b_f} &= \text{Var}_{b_{f_n}}[\theta]. \end{aligned} \quad (11.47)$$

According to the EP update rule, the message from factor f_n to θ is

$$\Delta_n^{f \rightarrow \theta}(\theta) = \mathcal{UN}(\theta | \mu_n^{f \rightarrow \theta}, \tau_n^{f \rightarrow \theta}) \propto \frac{\mathcal{UN}(\theta | \mu_n^{b_f}, \tau_n^{b_f})}{\Delta_n^{f \leftarrow \theta}(\theta)}, \quad (11.48)$$

where

$$\begin{aligned} \tau_n^{f \rightarrow \theta} &= \left(\frac{1}{\tau_n^{b_f}} - \frac{1}{\tau_n^{f \leftarrow \theta}} \right)^{-1} \\ \mu_n^{f \rightarrow \theta} &= \tau_n^{f \rightarrow \theta} \left(\frac{\mu_n^{b_f}}{\tau_n^{b_f}} - \frac{\mu_n^{f \leftarrow \theta}}{\tau_n^{f \leftarrow \theta}} \right). \end{aligned} \quad (11.49)$$

After the final iteration, the Gaussian approximation of the univariate factored PDF is the belief at variable θ :

$$b_\theta(\theta) \propto \prod_n \Delta_n^{f \rightarrow \theta}(\theta). \quad (11.50)$$

Therefore, the mean and variance of $b_\theta(\theta)$ are the approximated mean and variance of the univariate factored PDF:

$$\begin{aligned} \tau^{b_\theta} &= \left(\sum_n \frac{1}{\tau_n^{f \rightarrow \theta}} \right)^{-1} \\ \mu^{b_\theta} &= \tau^{b_\theta} \left(\sum_n \frac{\mu_n^{f \rightarrow \theta}}{\tau_n^{f \rightarrow \theta}} \right). \end{aligned} \quad (11.51)$$

The above procedure is then repeated for the next factor f_k , where $k = (n \bmod N) + 1$. Since a sequential update strategy is used, $b_\theta(\theta)$ is updated by (11.50) whenever a factor-to-variable message is updated. Thus, (11.50) serves as the definition for $b_\theta(\theta)$ throughout the algorithm.

Proposition 1. *Assume that $b_\theta(\theta)$ is defined as in (11.50). If the algorithm in (11.44)-(11.51) is initialized with messages such that $b_\theta(\theta)$ is integrable, then the integral of $b_\theta(\theta)$ remains finite throughout all iterations.*

Proof. We prove the proposition by mathematical induction over the sub-iterations. The i -th sub-iteration is defined as the update in which only the messages to and from the n -th factor are considered, where $n = [(i-1) \bmod N] + 1$. An iteration therefore consists of N sub-iterations.

By assumption, the belief $b_\theta(\theta)$ is initialized to be integrable.

For clarity, we denote $\forall k = \{1, \dots, N\} : \Delta_{k;(i)}^{f \leftarrow \theta}(\theta)$ and $\Delta_{k;(i)}^{f \rightarrow \theta}(\theta)$ as the messages to and from factor $f_k(\theta)$ at the end of sub-iteration i . Similarly, let $b_{\theta;(i)}(\theta)$ denote the variable-level belief at the end of sub-iteration i . We also denote $\mu_{n;(i)}^{b_f}$ and $\tau_{n;(i)}^{b_f}$ as the mean and variance of the factor-level belief $b_{f_n}(\theta)$ at the end of iteration i .

Assume that at the end of sub-iteration i , the belief $b_{\theta;(i)}(\theta)$ is integrable. At sub-iteration $i+1$, the messages to and from factor $f_m(\theta)$ are updated, where $m = (i \bmod N) + 1$. If the factor-level belief computed in (11.46) is not integrable, i.e.,

$$\int f_m(\theta) \Delta_{m;(i+1)}^{f \leftarrow \theta}(\theta) d\theta = \infty, \quad (11.52)$$

then the update of $\Delta_m^{f \rightarrow \theta}(\theta)$ is skipped and

$$\Delta_{m;(i+1)}^{f \rightarrow \theta}(\theta) = \Delta_{m;(i)}^{f \rightarrow \theta}(\theta). \quad (11.53)$$

Consequently,

$$b_{\theta;(i+1)}(\theta) = b_{\theta;(i)}(\theta), \quad (11.54)$$

which remains integrable.

If instead the factor-level belief computed in (11.46) is integrable, i.e.,

$$\int f_m(\theta) \Delta_{m;(i+1)}^{f \leftarrow \theta}(\theta) d\theta < \infty, \quad (11.55)$$

Algorithm 11.2: Persistent EP for GMM Factors

```

1: Initialize  $\forall n : \mu_n^{f \rightarrow \theta} = 0, \tau_n^{f \rightarrow \theta} = 1$ 
2: repeat
3:   [For every  $n \in [1, \dots, N]$ , repeat the following]
4:   Update  $\tau^{b_\theta}$  and  $\mu^{b_\theta}$  by (11.51)
5:    $\tau_n^{f \leftarrow \theta} = \frac{\tau_n^{f \rightarrow \theta} \tau^{b_\theta}}{\tau_n^{f \rightarrow \theta} - \tau^{b_\theta}}$ 
6:    $\mu_n^{f \leftarrow \theta} = \frac{\tau_n^{f \rightarrow \theta} \mu^{b_\theta} - \tau^{b_\theta} \mu_n^{f \rightarrow \theta}}{\tau_n^{f \rightarrow \theta} - \tau^{b_\theta}}$ 
7:   if  $b_{f_n}(\theta)$  in (11.46) is integrable then
8:     Update  $\mu_n^{b_f}$  and  $\tau_n^{b_f}$  by (11.46) - (11.47)
9:     Update  $\tau_n^{f \rightarrow \theta}$  and  $\mu_n^{f \rightarrow \theta}$  by (11.49)
10:  end if
11: until Convergence
12: Output  $\mu^{b_\theta}$  and  $\tau^{b_\theta}$  as approximated mean and variance

```

then by (11.48), we have

$$\Delta_{m;(i+1)}^{f \rightarrow \theta}(\theta) \Delta_{m;(i+1)}^{f \leftarrow \theta}(\theta) = \mathcal{UN}(\theta | \mu_{m;(i+1)}^{b_f}, \tau_{m;(i+1)}^{b_f}), \quad (11.56)$$

where $\tau_{m;(i+1)}^{b_f} \geq 0$. Since $\tau_{m;(i+1)}^{b_f} \geq 0$, we can verify (11.56) has finite integral (if $\tau_{m;(i+1)}^{b_f} = 0$, then (11.56) is Dirac delta function).

Finally, comparing (11.44) and (11.50), we have

$$b_{\theta;(i+1)}(\theta) = \Delta_{m;(i+1)}^{f \rightarrow \theta}(\theta) \Delta_{m;(i+1)}^{f \leftarrow \theta}(\theta). \quad (11.57)$$

Thus, $b_{\theta;(i+1)}(\theta)$ is also integrable.

By induction, $b_\theta(\theta)$ has a finite integral at every sub-iteration. $\square$

The persistent EP algorithm for GMM factors is concluded in Algorithm 11.2. Due to the sequential update strategy, line 4 of Algorithm 11.2 has computational complexity $O(1)$, since only a single term corresponding to the most recently updated message $\Delta_k^{f \rightarrow \theta}$ is updated when updating $b_\theta(\theta)$, where $k = [(n - 2) \bmod N] + 1$.

11.6 Analytic Continuation EP

To avoid non-integrable beliefs that prevent the algorithm from progressing, we adopt the technique of analytic continuation proposed in [70]. While persistent EP handles non-integrable beliefs in a passive manner, Analytic Continuation EP (ACEP) actively prevents non-integrable factor-level belief at the subsequent factor. By expressing Gaussian distributions in terms of their natural parameters as in (11.40), we have a removable discontinuity at $\xi = 0$ in $\mathcal{UM}$ if we discard the constant term $-\frac{\nu^2}{\xi}$, and the essential discontinuity $\tau = 0$ in $\mathcal{UN}$ becomes $\xi \rightarrow \pm\infty$. Consequently, it is more convenient in this section to represent Gaussian distributions using the natural parameters (ν, ξ) rather than the mean-variance pair (μ, τ) .

Based on EP, the message from θ to the n -th factor is updated as

$$\Delta_n^{f \leftarrow \theta}(\theta) = \mathcal{UM}(\theta | \nu_n^{f \leftarrow \theta}, \xi_n^{f \leftarrow \theta}) \propto \prod_{n' \neq n} \Delta_{n'}^{f \rightarrow \theta}(\theta), \quad (11.58)$$

where

$$\begin{aligned}\nu_n^{f \leftarrow \theta} &= \sum_{n' \neq n} \nu_{n'}^{f \rightarrow \theta}; \\ \xi_n^{f \leftarrow \theta} &= \sum_{n' \neq n} \xi_{n'}^{f \rightarrow \theta}.\end{aligned}\tag{11.59}$$

The belief at f_n is then computed as

$$b_{f_n}(\theta) \propto f_n(\theta) \Delta_n^{f \leftarrow \theta}(\theta).\tag{11.60}$$

When the Gaussian family is chosen as the projection family in EP, minimizing the Kullback–Leibler divergence is equivalent to matching the mean and variance. Accordingly, we define the approximated posterior mean and variance of the belief at factor f_n as

$$\begin{aligned}\mu_n^{b_f} &= \mathbb{E}_{b_{f_n}}[\theta], \\ \tau_n^{b_f} &= \text{Var}_{b_{f_n}}[\theta].\end{aligned}\tag{11.61}$$

Following (11.61), the corresponding natural parameters are defined as

$$\begin{aligned}\nu_n^{b_f} &= \frac{\mu_n^{b_f}}{\tau_n^{b_f}}; \\ \xi_n^{b_f} &= \frac{1}{\tau_n^{b_f}}.\end{aligned}\tag{11.62}$$

Inspired by [70], to prevent non-integrable beliefs from blocking further updates, it suffices to ensure that the belief at the next factor $f_k(\theta)$ remains integrable, where $k = (n \bmod N) + 1$.

Due to the following two observations:

- $(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta})$ and $(\nu_k^{f \leftarrow \theta}, \xi_k^{f \leftarrow \theta})$ are linearly related through (11.59);
- no other messages are updated between the updates of $\Delta_n^{f \rightarrow \theta}(\theta)$ and $\Delta_k^{f \leftarrow \theta}(\theta)$,

and hence, there exists a bijection between $(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta})$ and $(\nu_k^{f \leftarrow \theta}, \xi_k^{f \leftarrow \theta})$. We therefore define a domain Ω_n for $(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta})$ such that, if $(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) \in \Omega_n$ the belief update at the next factor node

$$b_{f_k}(\theta) \propto f_k(\theta) \Delta_k^{f \leftarrow \theta}(\theta)\tag{11.63}$$

is integrable, where $k = (n \bmod N) + 1$.

Based on the EP rules for marginal factor-level belief projection and feedback message, ACEP merges the projection and message-update steps into a direct update of the factor-to-variable message.

Due to the feedback message update rule, the projected belief admits the form

$$\begin{aligned}\hat{b}_{f_n}(\theta) &= \mathcal{M} \left[\theta \left| \nu_n^{\hat{b}_f}(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}), \xi_n^{\hat{b}_f}(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) \right. \right] \\ &\propto \Delta_n^{f \leftarrow \theta}(\theta) \Delta_n^{f \rightarrow \theta}(\theta),\end{aligned}\tag{11.64}$$

with

$$\begin{aligned}\nu_n^{\hat{b}_f}(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) &= \nu_n^{f \rightarrow \theta} + \nu_n^{f \leftarrow \theta} \\ \xi_n^{\hat{b}_f}(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) &= \xi_n^{f \rightarrow \theta} + \xi_n^{f \leftarrow \theta}.\end{aligned}\tag{11.65}$$

Since ACEP is iterative, the parameters $(\nu_n^{f \leftarrow \theta}, \xi_n^{f \leftarrow \theta})$ are treated as constants during the update of $\Delta_n^{f \rightarrow \theta}(\theta)$. Substituting (11.64) and (11.65) into the KLD projection as the second KLD parameter, and incorporating the domain constraint Ω_n , yields

$$(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) = \arg \min_{(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) \in \Omega_n} \text{KLD} \left[b_{f_n}(\theta) \| \hat{b}_{f_n}(\theta) \right], \quad (11.66)$$

where the left KLD parameter $b_{f_n}(\theta)$ is defined in (11.60). Expanding the KLD and discarding the constants gives

$$(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) = \arg \min_{(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) \in \Omega_n} L_n(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}), \quad (11.67)$$

where

$$\begin{aligned} L_n(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) = & -\frac{\ln \left[\xi_n^{\hat{b}_f}(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) \right]}{2} \\ & + \frac{\left[\nu_n^{\hat{b}_f}(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) - \xi_n^{\hat{b}_f}(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) \mu_n^{b_f} \right]^2}{2 \xi_n^{\hat{b}_f}(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta})} \\ & + \frac{\xi_n^{\hat{b}_f}(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta}) \tau_n^{b_f}}{2}. \end{aligned} \quad (11.68)$$

For the unconstrained case $\Omega_n = \mathbb{R}^2$, the optimal point of (11.67) is obtained by setting the partial derivative with respect to each of the parameters to zero:

$$\begin{aligned} \xi_n^{f \rightarrow \theta} &= \frac{1}{\tau_n^{b_f}} - \xi_n^{f \leftarrow \theta} \\ \nu_n^{f \rightarrow \theta} &= (\xi_n^{f \rightarrow \theta} + \xi_n^{f \leftarrow \theta}) \mu_n^{b_f} - \nu_n^{f \leftarrow \theta} \\ &= \frac{\mu_n^{b_f}}{\tau_n^{b_f}} - \nu_n^{f \leftarrow \theta}. \end{aligned} \quad (11.69)$$

Compared to the mean-variance representation, the natural-parameter representation allows smooth transition as $\xi_n^{f \rightarrow \theta}$ crosses zero. Thus, $\xi_n^{f \rightarrow \theta}$ and $\nu_n^{f \rightarrow \theta}$ remain well-defined in $\mathbb{R}$ as long as the belief $b_{f_n}(\theta)$ is integrable with well-defined mean and variance.

Finally, we approximate the mean and variance of the given univariate factored PDF using those of the variable-level belief:

$$b_\theta(\theta) = \mathcal{M}(\theta | \nu^{b_\theta}, \xi^{b_\theta}) \propto \prod_n \Delta_n^{f \rightarrow \theta}(\theta), \quad (11.70)$$

with

$$\begin{aligned} \nu^{b_\theta} &= \sum_n \nu_n^{f \rightarrow \theta} \\ \xi^{b_\theta} &= \sum_n \xi_n^{f \rightarrow \theta}. \end{aligned} \quad (11.71)$$

The corresponding belief variance and mean are

$$\begin{aligned} \tau^{b_\theta} &= \frac{1}{\xi^{b_\theta}} \\ \mu^{b_\theta} &= \frac{\nu^{b_\theta}}{\xi^{b_\theta}}. \end{aligned} \quad (11.72)$$

Proposition 2. Assume that $b_\theta(\theta)$ is defined as in (11.70). If the algorithm in (11.58)-(11.72) is initialized with messages such that $b_\theta(\theta)$ is integrable, then it remains integrable at every iteration.

Proof. As in Proposition 1, we proceed by mathematical induction. By assumption, the initial $b_\theta(\theta)$ is integrable.

As the induction hypothesis, assume that after updating the message $\Delta_k^{f \rightarrow \theta}(\theta)$, the belief $b_\theta(\theta)$ remains integrable.

Let $n = (k \bmod N) + 1$. The algorithm enforces that the factor-level belief $b_{f_n}(\theta)$ is integrable. Consequently, the objective function (11.68) is well defined. From the first term on the right-hand side of (11.68), we see that, for L_n to be well defined, it is required that

$$\xi_n^{f \rightarrow \theta} + \xi_n^{f \leftarrow \theta} > 0, \quad (11.73)$$

which defines the natural domain for L_n .

After updating $\xi_n^{f \rightarrow \theta}$, the precision (inverted variance) of the variable-level belief becomes

$$\xi^{b_\theta} = \xi_n^{f \rightarrow \theta} + \xi_n^{f \leftarrow \theta} > 0. \quad (11.74)$$

Therefore, the belief $b_\theta(\theta)$ remains integrable. $\square$

11.6.1 GMM Case

As an illustrative example, consider Gaussian mixture factors of the form given in (11.26)

$$f_n(\theta) \propto \sum_{s_n} p_{s_n} \mathcal{M}(\theta | \nu_{s_n}, \xi_{s_n}), \quad (11.75)$$

where

$$\begin{aligned} \nu_{s_n} &= \frac{\mu_{s_n}}{\tau_{s_n}}; \\ \xi_{s_n} &= \frac{1}{\tau_{s_n}}. \end{aligned} \quad (11.76)$$

From (11.60) and (11.42), the factor-level belief at f_n can be written as

$$b_{f_n}(\theta) \propto \sum_{s_n} \omega_{s_n} \mathcal{M}(\theta | \nu_{s_n} + \nu_n^{f \leftarrow \theta}, \xi_{s_n} + \xi_n^{f \leftarrow \theta}), \quad (11.77)$$

where

$$\omega_{s_n} = \frac{p_{s_n} \sqrt{\xi_{s_n}}}{\sqrt{\xi_{s_n} + \xi_n^{f \leftarrow \theta}}} \mathcal{U} \mathcal{M} \left(0 \left| \frac{\xi_n^{f \leftarrow \theta} \nu_{s_n} - \xi_{s_n} \nu_n^{f \leftarrow \theta}}{\xi_{s_n} + \xi_n^{f \leftarrow \theta}}, \frac{\xi_{s_n} \xi_n^{f \leftarrow \theta}}{\xi_{s_n} + \xi_n^{f \leftarrow \theta}} \right. \right). \quad (11.78)$$

As discussed earlier, $\xi_n^{f \leftarrow \theta}$ may be zero, which is a discontinuity in (11.78). However, due to the design of the additional domain Ω_n as we will see in (11.83) and (11.84), we use the ansatz that $\forall s_n : \xi_{s_n} + \xi_n^{f \leftarrow \theta} \geq 0$. To analyze the asymptotic behavior of (11.77), we

examine the ratio between $\omega_{s'_n}$ and ω_{s_n}

$$\begin{aligned} \rho_{s_n}^{s'_n} &= \frac{\omega_{s'_n}}{\omega_{s_n}} = \frac{p_{s'_n} \sqrt{\xi_{s'_n} (\xi_{s_n} + \xi_n^{f \leftarrow \theta})}}{p_{s_n} \sqrt{\xi_{s_n} (\xi_{s'_n} + \xi_n^{f \leftarrow \theta})}} \\ &\cdot \exp \left(-\frac{1}{2} \left[\frac{\xi_n^{f \leftarrow \theta} \nu_{s'_n}^2}{\xi_{s'_n} (\xi_{s'_n} + \xi_n^{f \leftarrow \theta})} - \frac{2\nu_{s'_n} \nu_n^{f \leftarrow \theta}}{\xi_{s'_n} + \xi_n^{f \leftarrow \theta}} + \frac{2\nu_{s_n} \nu_n^{f \leftarrow \theta}}{\xi_{s_n} + \xi_n^{f \leftarrow \theta}} \right. \right. \\ &\quad \left. \left. - \frac{\xi_n^{f \leftarrow \theta} \nu_{s_n}^2}{\xi_{s_n} (\xi_{s_n} + \xi_n^{f \leftarrow \theta})} + \frac{\nu_n^{f \leftarrow \theta}}{\xi_{s_n} + \xi_n^{f \leftarrow \theta}} - \frac{\nu_n^{f \leftarrow \theta}}{\xi_{s'_n} + \xi_n^{f \leftarrow \theta}} \right] \right). \end{aligned} \quad (11.79)$$

From (11.79), we observe that $\rho_{s_n}^{s'_n}$ remains finite and positive if $\xi_n^{f \leftarrow \theta} = 0$. If $\xi_n^{f \leftarrow \theta} \rightarrow (-\xi_{s'_n})^+$, we have $\rho_{s_n}^{s'_n} \rightarrow 0$. Moreover, we can verify that if $\xi_n^{f \leftarrow \theta} \rightarrow (-\xi_{s'_n})^+$, $\rho_{s_n}^{s'_n}$ is a higher order infinitesimal of $\xi_n^{f \leftarrow \theta} + \xi_{s'_n}$. A similar asymptotic behavior at $\xi_n^{f \leftarrow \theta} \rightarrow (-\xi_{s_n})^+$ can be obtained by analyzing $\rho_{s'_n}^{s_n}$. Consequently, (11.77) can be equivalently written as

$$\forall s_n : b_{f_n}(\theta) \propto \sum_{s'_n} \rho_{s_n}^{s'_n} \mathcal{M}(\theta | \nu_{s'_n} + \nu_n^{f \leftarrow \theta}, \xi_{s'_n} + \xi_n^{f \leftarrow \theta}). \quad (11.80)$$

To normalize (11.80), we define the normalized ratio

$$\bar{\rho}_{s_n}^{s'_n} = \frac{\rho_{s_n}^{s'_n}}{\sum_{s'_n} \rho_{s_n}^{s'_n}}. \quad (11.81)$$

The mean and variance of $b_{f_n}(\theta)$ in (11.80) are given by

$$\begin{aligned} \mu_n^{b_f} &= \sum_{s'_n} \frac{\bar{\rho}_{s_n}^{s'_n} (\nu_{s'_n} + \nu_n^{f \leftarrow \theta})}{\xi_{s'_n} + \xi_n^{f \leftarrow \theta}} \\ \tau_n^{b_f} &= \left(\sum_{s'_n} \bar{\rho}_{s_n}^{s'_n} \left[\frac{1}{\xi_{s'_n} + \xi_n^{f \leftarrow \theta}} + \left(\frac{\nu_{s'_n} + \nu_n^{f \leftarrow \theta}}{\xi_{s'_n} + \xi_n^{f \leftarrow \theta}} \right)^2 \right] \right) - (\mu_n^{b_f})^2. \end{aligned} \quad (11.82)$$

To close the loop, we update the message $\Delta_n^{f \rightarrow \theta}(\theta)$ using (11.66) - (11.68). Let the next factor be $k = (n \bmod N) + 1$. Since the integrability of a GMM is determined by the signs of its component variances/precisions, we define the threshold for $\xi_n^{f \rightarrow \theta}$ as

$$\begin{aligned} \xi_n^{\text{thres}} &= -\min_{s_k} (\xi_{s_k}) - \prod_{n' \neq n, n' \neq k} \xi_{n'}^{f \rightarrow \theta} \\ &= -\min_{s_k} (\xi_{s_k}) - \xi_n^{f \leftarrow \theta} + \xi_k^{f \rightarrow \theta}. \end{aligned} \quad (11.83)$$

Thus, the admissible domain of $(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta})$ that ensures the integrability of $b_{f_k}(\theta)$ is

$$\Omega_n = \mathbb{R} \times (\xi_n^{\text{thres}}, +\infty), \quad (11.84)$$

where $(\xi_n^{\text{thres}}, \infty)$ denotes an open interval from ξ_n^{thres} to $+\infty$. However, we can include ξ_n^{thres} by using the asymptotic analysis above.

To solve the KLD projection in (11.66), since $\nu_n^{f \rightarrow \theta}$ is unconstrained, we first set the partial derivative with respect to $\nu_n^{f \rightarrow \theta}$ to zero, which gives us

$$\nu_n^{f \rightarrow \theta} = (\xi_n^{f \rightarrow \theta} + \xi_n^{f \leftarrow \theta}) \mu_n^{b_f} - \nu_n^{f \leftarrow \theta}. \quad (11.85)$$

Substituting (11.85) into (11.68), we can verify that $L_n(\nu_n^{f \rightarrow \theta}, \xi_n^{f \rightarrow \theta})$ is decreasing for

$$\xi_n^{f \rightarrow \theta} \in \left(-\xi_n^{f \leftarrow \theta}, \frac{1}{\tau_n^{b_f}} - \xi_n^{f \leftarrow \theta} \right], \quad (11.86)$$

and increasing for

$$\xi_n^{f \rightarrow \theta} \in \left(\frac{1}{\tau_n^{b_f}} - \xi_n^{f \leftarrow \theta}, +\infty \right). \quad (11.87)$$

Therefore, $\xi_n^{f \rightarrow \theta}$ is obtained as

$$\xi_n^{f \rightarrow \theta} = \begin{cases} \frac{1}{\tau_n^{b_f}} - \xi_n^{f \leftarrow \theta}, & \text{if } \frac{1}{\tau_n^{b_f}} - \xi_n^{f \leftarrow \theta} > \xi_n^{\text{thres}} \\ \xi_n^{\text{thres}}, & \text{otherwise} \end{cases}. \quad (11.88)$$

By substituting (11.88) back into (11.85), we obtain the update for the complete message $\Delta_n^{f \rightarrow \theta}(\theta)$.

After the final iteration, the mean and variance of the product of GMM factors can be obtained from the variable-level belief

$$b_\theta(\theta) = \prod_n \Delta_n^{f \rightarrow \theta}(\theta). \quad (11.89)$$

Accordingly, the mean and variance of $b_\theta(\theta)$ can be explicitly computed as

$$\begin{aligned} \mu^{b_\theta} &= \frac{\sum_n \nu_n^{f \rightarrow \theta}}{\sum_n \xi_n^{f \rightarrow \theta}}; \\ \tau^{b_\theta} &= \frac{1}{\sum_n \xi_n^{f \rightarrow \theta}}. \end{aligned} \quad (11.90)$$

This concludes the ACEP algorithm for Gaussian mixture model factors, as summarized in Algorithm 11.3.

11.7 Simulation Results

To verify the effectiveness of the proposed methods, we generate factored PDFs consisting of $N = 8$ GMM factors in each realization. Each GMM factor contains two Gaussian components with randomly assigned weights, means, and variances. The product of these 8 GMM factors results in another GMM with 2^8 components. To assess the approximation accuracy, we compute the exact mean and variance of the resulting 2^8 -component GMM using brute force, denoted by $\mu_{GMM}^{\hat{\theta}}$ and $\tau_{GMM}^{\hat{\theta}}$ respectively.

The normalized squared error (NSE) for a single realization is defined as

$$\begin{aligned} NSE_\mu &= \frac{(\mu^{\hat{\theta}} - \mu_{GMM}^{\hat{\theta}})^2}{(\mu_{GMM}^{\hat{\theta}})^2} \\ NSE_\tau &= \frac{(\tau^{\hat{\theta}} - \tau_{GMM}^{\hat{\theta}})^2}{(\tau_{GMM}^{\hat{\theta}})^2}. \end{aligned} \quad (11.91)$$

Algorithm 11.3: ACEP for GMM Factors

```

1: Initialize  $\forall n : \nu_n^{f \rightarrow \theta} = 0, \xi_n^{f \rightarrow \theta} = 1$ 
2: repeat
3:   [For every  $n \in [1, \dots, N]$ , repeat the following]
4:   Update  $\nu^{b_\theta}$  and  $\xi^{b_\theta}$  by (11.71)
5:    $\xi_n^{f \leftarrow \theta} = \xi^{b_\theta} - \xi_n^{f \rightarrow \theta}$ 
6:    $\nu_n^{f \leftarrow \theta} = \nu^{b_\theta} - \nu_n^{f \rightarrow \theta}$ 
7:   Choose the index of a GMM component  $s_n$ .
8:    $\forall s'_n$  : compute  $\rho_{s'_n}^{s'_n}$  by (11.79)
9:    $\forall s'_n$  : compute  $\bar{\rho}_{s'_n}^{s'_n}$  by (11.81)
10:  Update  $\mu_n^{b_f}$  and  $\tau_n^{b_f}$  by (11.82)
11:   $k = (n \bmod N) + 1$ 
12:  Update  $\xi_n^{\text{thres}}$  by (11.83)
13:  Update  $\xi_n^{f \rightarrow \theta}$  by (11.88)
14:  Update  $\nu_n^{f \rightarrow \theta}$  by (11.85)
15: until Convergence
16: Output  $\mu^{b_\theta}$  and  $\tau^{b_\theta}$  as mean and variance by (11.90)

```

We simulated 10,000 realizations with different sets of GMM factors. The proposed methods are compared with the commonly used Clipping EP [2, 19]. Since checking the integrability of the belief $b_{f_n}(\theta)$ can be computationally expensive, we consider two variants for both proposed EP-based algorithms:

- **Strict approach (STR.):** the algorithms are implemented exactly as proposed.
- **Relaxed approach (RLX.):** for persistent EP, we check the integrability of $\Delta_n^{f \leftarrow \theta}(\theta)$ instead of $b_{f_n}(\theta)$. For ACEP, we simply set ξ_n^{thres} to zero.

The simulation results for mean estimations are shown in Fig. 11.3, while the results for variance estimation are shown in Fig. 11.4. The abbreviation “PEP” denotes persistent EP.

In both figures, the curves corresponding to strict persistent EP, relaxed persistent EP, and strict ACEP coincide and consistently exhibit the best performance. All proposed methods outperform clipping EP in terms of NSE at the 95% CDF.

An interesting observation is that the mean estimate produced by relaxed ACEP is close to the best-performing methods, whereas its variance estimate is close to that of clipping EP. This behavior can be explained by comparing relaxed ACEP with strict ACEP and clipping EP. The variance estimation is closely related to the natural parameter ξ . Both relaxed ACEP and clipping EP disallow non-integrable messages. Therefore, Fig. 11.4 suggests that the optimal set of messages may include non-integrable messages. By examining the updates of the mean and natural parameter ν in clipping EP and ACEP, we observe that ACEP exploits the limit around the discontinuity $\xi_n^{f \rightarrow \theta} = 0$. In contrast, in clipping EP, when $\tau_n^{f \rightarrow \theta}$ is clipped to $+\infty$, $\mu_n^{f \rightarrow \theta}$ is still treated as a finite value. However, as indicated by ACEP, $\mu_n^{f \rightarrow \theta}$ should instead scale as $O(\tau_n^{f \rightarrow \theta})$.

11.8 Conclusions

This paper proposes one BP-based algorithm and two EP-based algorithms for estimating the means and variances of univariate factored PDFs.

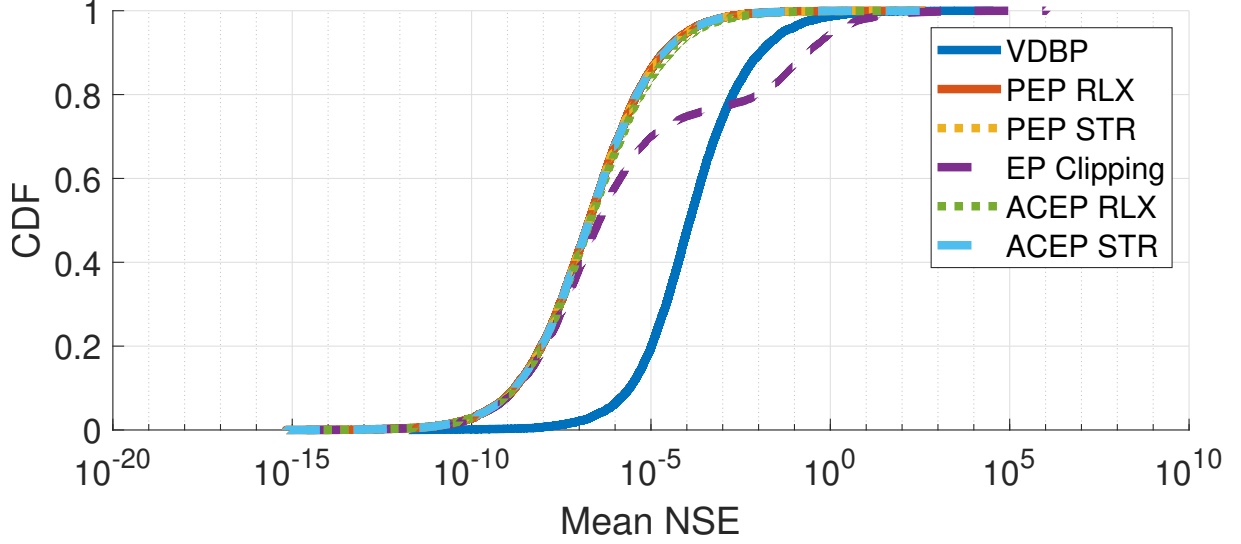


Figure 11.3: NSE CDF of mean estimates.

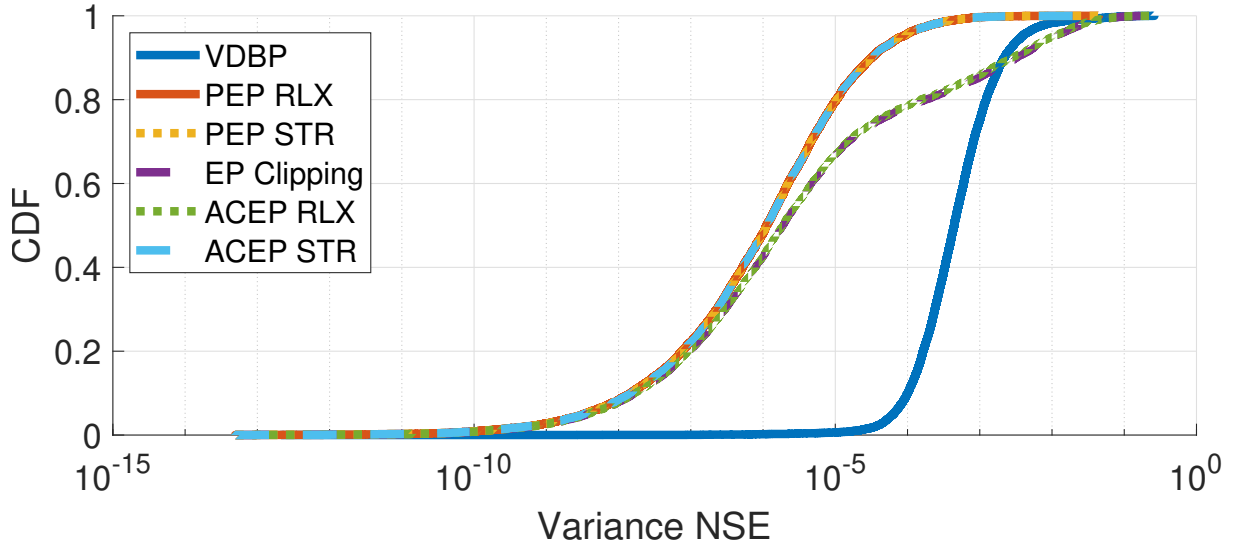


Figure 11.4: NSE CDF of variance estimates.

In the BP-based method, VDBP, we construct a multivariate measurement model whose posterior can be marginalized to the given univariate factored PDF. GaBP is then applied to this multivariate measurement model to approximate the marginal posterior.

EP can be directly applied to estimate the mean and variance of a univariate factored PDF when all factor-level beliefs remain integrable throughout the iterations. To address scenarios involving non-integrable factor-level beliefs, we proposed persistent EP and ACEP.

Persistent EP adopts a passive strategy by skipping message updates around a factor whenever the corresponding belief at the factor is non-integrable. In contrast, ACEP takes an active approach by modifying the projection family to ensure that the belief at the next factor is always integrable.

Chapter 12

Avoiding Non-Integrable Beliefs in Expectation Propagation

12.1 Introduction

Signal recovery is a fundamental problem in signal processing, with a wide range of applications. In the Bayesian framework, however, canonical inference methods such as MMSE and MAP suffer from exponential computational complexity as the problem dimension increases.

Graphical-model-based iterative methods have proven effective by exploiting structural properties of the underlying models [5]. Expectation Propagation (EP), introduced in [35], transforms a global inference problem into multiple local inference problems. EP can be interpreted as an iterative procedure for finding the stationary point of the constrained Bethe Free Energy (BFE) [6, 11]. Within this framework, EP approximates marginal distributions (beliefs) using functions known as messages, which correspond one-to-one with the Lagrange multipliers in the constrained BFE formulation. The iterative message passing is interpreted as the alternating optimization of the corresponding constrained BFE.

In EP and constrained BFE minimization, beliefs are required to be integrable. Whereas, the messages are allowed to have any form. However, the alternating optimization or EP cannot foresee whether a message update will lead to non-integrable belief.

In most literature [10, 19, 28], the authors restrict the algorithm to integrable messages to avoid non-integrable beliefs. However, such augmentations are suboptimal, since optimal solution may contain non-integrable messages [70, 72].

12.1.1 Main Contributions

In [70], we investigate the non-integrable beliefs and message in linear model. However, the proposed methods only work for simple factor graph and cannot be extended to more general ones such as Generalized Linear Model (GLM) or bilinear estimation problems. Furthermore, the analytical continuation technique was only used to restrict the messages to be integrable.

In this chapter, we extend the ideas in [70] and propose two frameworks for EP to avoid non-integrable belief. Both of the methods requires sequential update. GLM is then investigated using our proposed EP frameworks.

For readers who is interested, please find all the proofs in [73].

12.2 Analytic Continuation Expectation Propagation (ACEP)

Consider a factored distribution

$$p(\boldsymbol{\theta}) \propto \prod_{\alpha} f_{\alpha}(\boldsymbol{\theta}_{\alpha}). \quad (12.1)$$

We adopt a sequential update scheme to infer the marginal distribution of $p(\boldsymbol{\theta})$ as $b_n^{\hat{\theta}}(\theta_n) \simeq \int p(\boldsymbol{\theta}) \prod_{n' \neq n} d\theta_{n'}$. The belief at factor f_{α} can be computed as

$$b_{\alpha}^f(\boldsymbol{\theta}_{\alpha}) \propto f_{\alpha}(\boldsymbol{\theta}_{\alpha}) \prod_{n \in N(\alpha)} \mu_{\alpha,n}^{\theta \rightarrow f}(\theta_n) \quad (12.2)$$

The marginal belief is

$$b_{\alpha,n}^{\hat{\theta};f}(\theta_n) = \int b_{\alpha}^f(\boldsymbol{\theta}_{\alpha}) \prod_{n' \in N(\alpha)/\{n\}} d\theta_{n'}. \quad (12.3)$$

It is then projected to Gaussian distribution by KLD

$$\min_{\mu_{\alpha,n}^{f \rightarrow \theta}(\theta_n) \in \mathcal{F}_n} \text{KLD}[b_{\alpha,n}^{\hat{\theta};f}(\theta_n) \parallel \frac{1}{Z} \mu_{\alpha,n}^{f \rightarrow \theta}(\theta_n) \mu_{\alpha,n}^{\theta \rightarrow f}(\theta_n)], \quad (12.4)$$

where $\mathcal{F}_n$ ensures that $b_n^{\hat{\theta}}(\theta_n)$ is integrable.

The belief at the variable node is

$$b_n^{\hat{\theta}}(\theta_n) \propto \prod_{\alpha \in N(n)} \mu_{\alpha,n}^{f \rightarrow \theta}(\theta_n) \quad (12.5)$$

It is then projected to Gaussian by KLD

$$\min_{\mu_{\alpha,n}^{\theta \rightarrow f}(\theta_n) \in \mathcal{F}_{\alpha}} \text{KLD}[b_n^{\hat{\theta}}(\theta_n) \parallel \frac{1}{Z} \mu_{\alpha,n}^{f \rightarrow \theta}(\theta_n) \mu_{\alpha,n}^{\theta \rightarrow f}(\theta_n)], \quad (12.6)$$

where the domain $\mathcal{F}_{\alpha}$ ensures that the belief $b_{\alpha}^f(\boldsymbol{\theta}_{\alpha})$ is integrable.

12.2.1 Constrained KLD Optimization in Gaussian Projection

Gaussian projection is commonly used in EP. When using Gaussian projection, a common problem is “negative variance”, which may lead to non-integrable belief. non-integrable messages are acceptable, while non-integrable belief prevents the algorithm from proceeding. We define unnormalized Gaussian

$$\overline{\mathcal{N}}(\boldsymbol{\theta} | \mathbf{m}, \mathbf{C}) = \exp[-(\boldsymbol{\theta} - \mathbf{m})^H \mathbf{C}^{-1} (\boldsymbol{\theta} - \mathbf{m})]. \quad (12.7)$$

Sometimes, it is more convenient to use natural parameters, we define unnormalized Gaussian with natural parameter to be

$$\overline{\mathcal{M}}(\boldsymbol{\theta} | \boldsymbol{\nu}, \boldsymbol{\Xi}) = \exp[-(\boldsymbol{\theta} - \boldsymbol{\Xi}^{-1} \boldsymbol{\nu})^H \boldsymbol{\Xi} (\boldsymbol{\theta} - \boldsymbol{\Xi}^{-1} \boldsymbol{\nu})]. \quad (12.8)$$

Therefore, we have

$$[(\boldsymbol{\nu}, \boldsymbol{\Xi}) = (\mathbf{C}^{-1} \mathbf{m}, \mathbf{C}^{-1})] \Leftrightarrow [\overline{\mathcal{N}}(\boldsymbol{\theta} | \mathbf{m}, \mathbf{C}) = \overline{\mathcal{M}}(\boldsymbol{\theta} | \boldsymbol{\nu}, \boldsymbol{\Xi})] \quad (12.9)$$

We consider the constrained projection step

$$\begin{aligned} \min_{\mu_p(\theta)} \quad & \text{KLD}[b(\theta) \| \frac{1}{Z} \mu_r(\theta) \mu_p(\theta)] \\ \text{s.t.} \quad & \mu_p \in \mathcal{F}, \end{aligned} \quad (12.10)$$

where $\mathcal{F}$ is a subset of unnormalized Gaussian. We often only have constraint on the variance (or precision) and the above problem becomes

$$\begin{aligned} \min_{\nu_p, \xi_p} \quad & \text{KLD}[b(\theta) \| \frac{1}{Z(\nu_p, \xi_p)} \mu_r(\theta) \overline{\mathcal{M}}(\theta | \nu_p, \xi_p)] \\ \text{s.t.} \quad & \xi_p > \gamma, \end{aligned} \quad (12.11)$$

We define

$$m_p = \xi_p^{-1} \nu_p \quad (12.12)$$

and thus,

$$\overline{\mathcal{M}}(\theta | \nu_p, \xi_p) = \exp[-(\theta - m_p)^T \xi_p (\theta - m_p)]. \quad (12.13)$$

Denote

$$\begin{aligned} m_{\hat{\theta}} &= \mathbb{E}_b[\theta] \\ \tau_{\hat{\theta}} &= \text{Var}_b[\theta]. \end{aligned} \quad (12.14)$$

For simplicity, denote

$$D = 2 \cdot \text{KLD}[b(\theta) \| \frac{1}{Z(\nu_p, \xi_p)} \mu_r(\theta) \overline{\mathcal{M}}(\theta | \nu_p, \xi_p)] \quad (12.15)$$

to remove the constant scale. Expand the KLD in (12.11),

$$\begin{aligned} D &= -\log[\det(\xi_p + \xi_r)] + \log(2\pi) \\ &+ \text{tr}[(\xi_p + \xi_r) \tau_{\hat{\theta}}] + [m_{\hat{\theta}} - (\xi_p + \xi_r)^{-1}(\nu_p + \nu_r)]^T \\ &\quad \cdot (\xi_p + \xi_r) [m_{\hat{\theta}} - (\xi_p + \xi_r)^{-1}(\nu_p + \nu_r)] \end{aligned} \quad (12.16)$$

We set the derivative w.r.t. ν_p to zero, and we have

$$\begin{aligned} (\xi_p + \xi_r)^{-1}(\nu_p + \nu_r) &= m_{\hat{\theta}} \\ \Rightarrow \nu_p &= (\xi_p + \xi_r) m_{\hat{\theta}} - \nu_r \end{aligned} \quad (12.17)$$

Substitute the result into the KLD, and we have

$$\begin{aligned} D|_{\nu_p=(\xi_p+\xi_r)m_{\hat{\theta}}-\nu_r} &= -\log[\det(\xi_p + \xi_r)] + \log(2\pi) \\ &+ \text{tr}[(\xi_p + \xi_r) \tau_{\hat{\theta}}] \end{aligned} \quad (12.18)$$

We can verify that (12.18) is convex even in multi-variate (θ is vector, ξ_p , ξ_r , $\tau_{\hat{\theta}}$ are matrices) cases. In multi-variate EP, the constraint in (12.11) becomes an intersection of positive-definite constraints, which forms an intersection of cones. Thus, the feasible set is also convex. Therefore, the constrained optimization of (12.18) is a convex optimization problem. We will only consider univariate case here, since multivariate does not change the convexity.

Set the derivative of (12.18) w.r.t. ξ_p , and take the constraint into consideration

$$\xi_p = \max\left\{\frac{1}{\tau_{\hat{\theta}}} - \xi_r, \gamma\right\}. \quad (12.19)$$

After that, ν_p can be updated by substituting (12.19) into (12.17). Since updating one message will only affect the belief at either a single factor or variable node, this modification won't increase the complexity order and it ensures that all the beliefs are integrable during each step.

12.3 EP with Checking (EPC)

Here we do not check if the beliefs are integrable. Instead, we skip the update of the messages from the node if the belief at the node is not integrable.

The belief at factor f_α can be computed as

$$b_\alpha^f(\boldsymbol{\theta}_\alpha) \propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{n \in N(\alpha)} \mu_{\alpha,n}^{\theta \rightarrow f}(\theta_n) \quad (12.20)$$

If this belief is not integrable, we just skip the message update from this node. The marginal belief is

$$b_{\alpha,n}^{\hat{\theta};f}(\theta_n) = \int b_\alpha^f(\boldsymbol{\theta}_\alpha) \prod_{n' \in N(\alpha)/\{n\}} d\theta_{n'}. \quad (12.21)$$

It is then projected to Gaussian distribution by KLD

$$\min_{\mu_{\alpha,n}^{f \rightarrow \theta}(\theta_n) \in \overline{\mathcal{M}}} \text{KLD}[b_{\alpha,n}^{\hat{\theta};f}(\theta_n) \parallel \frac{1}{Z} \mu_{\alpha,n}^{f \rightarrow \theta}(\theta_n) \mu_{\alpha,n}^{\theta \rightarrow f}(\theta_n)], \quad (12.22)$$

where $\overline{\mathcal{M}}$ denote the unnormalized Gaussian.

The belief at the variable node is

$$b_n^{\hat{\theta}}(\theta_n) \propto \prod_{\alpha \in N(n)} \mu_{\alpha,n}^{f \rightarrow \theta}(\theta_n) \quad (12.23)$$

The message from the variable is updated as

$$\mu_{\alpha,n}^{\theta \rightarrow f}(\theta_n) \propto \frac{b_n^{\hat{\theta}}(\theta_n)}{\mu_{\alpha,n}^{f \rightarrow \theta}(\theta_n)}. \quad (12.24)$$

We can verify that the belief at variable node is always integrable. In fact, we can rewrite the belief at variable node as

$$b_n^{\hat{\theta}}(\theta_n) \propto \mu_{\alpha,n}^{\theta \rightarrow f}(\theta_n) \mu_{\alpha,n}^{f \rightarrow \theta}(\theta_n), \quad (12.25)$$

which coincide with the second parameter in (12.22). Since we use unnormalized Gaussian family in the KLD projection, after updating the message $\mu_{\alpha,n}^{f \rightarrow \theta}(\theta_n)$ based on (12.22), the belief computed by (12.25) is a Gaussian with the same mean and variance as $b_{\alpha,n}^{\theta;f}(\theta_n)$.

12.4 System Model for Generalized Linear Model

We consider a generalized linear model (GLM), the unknown input $\mathbf{x}$ follows

$$\mathbf{x} \sim p(\mathbf{x}) = \prod_{k=1}^K p(x_k). \quad (12.26)$$

The observed output follows

$$\mathbf{y} \sim p(\mathbf{y} | \mathbf{A}^T \mathbf{x}) = \prod_{n=1:N} p(y_n | \mathbf{a}_n^T \mathbf{x}). \quad (12.27)$$

We introduce an auxiliary variable $\mathbf{z} = \mathbf{A}^T \mathbf{x}$. The joint PDF can be written as

$$\begin{aligned} p(\mathbf{y}, \mathbf{z}, \mathbf{x}) &= p(\mathbf{y}|\mathbf{z})p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) \\ &= \prod_n f_{y_n}(z_n) \cdot f_{\mathbf{z}}(\mathbf{z}, \mathbf{x}) \cdot \prod_k f_{x_k}(x_k) \end{aligned} \quad (12.28)$$

where for simplicity, we denote

$$f_{y_n}(z_n) = p(y_n|z_n) \quad (12.29)$$

$$f_{\mathbf{z}}(\mathbf{z}, \mathbf{x}) = \delta(\mathbf{z} - \mathbf{A}\mathbf{x}) \quad (12.30)$$

$$f_{x_k}(x_k) = p(x_k) \quad (12.31)$$

12.5 Derivation for GLM

12.5.1 $f_{y_n} \rightarrow z_n$

The belief at f_{y_n} is

$$b_n^{f_y}(z_n) \propto f_{y_n}(z_n) \mu_n^{z \rightarrow f_y}(z_n). \quad (12.32)$$

In ACEP, the above belief guarantees to be integrable. Meanwhile in EPC, the update of messages from the above belief is skipped if the above belief is not integrable. We assume it is integrable. The mean and variance are computed as

$$\begin{aligned} m_n^{\hat{z}; f_y} &= \mathbb{E}_{b_n^{f_y}}[z_n] \\ \tau_n^{\hat{z}; f_y} &= \text{Var}_{b_n^{f_y}}[z_n]. \end{aligned} \quad (12.33)$$

Next, we compute the threshold precision $\gamma_n^{f_y \rightarrow z}$. We only need to compute the threshold in ACEP when $\xi_n^{z \rightarrow f_y}$ is updated with the threshold value. Since the belief at the variable z_n is required to be integrable, we have

$$\gamma_n^{f_y \rightarrow z} + \xi_n^{f_z \rightarrow z} > 0 \Rightarrow \gamma_n^{f_y \rightarrow z} > -\xi_n^{f_z \rightarrow z}. \quad (12.34)$$

Therefore, we set

$$\gamma_n^{f_y \rightarrow z} = -\xi_n^{f_z \rightarrow z} + \epsilon, \quad (12.35)$$

where ϵ is a small positive number. Depending on which method is used, we have different ways of updating the message precision:

- If ACEP is used, we have

$$\xi_n^{f_y \rightarrow z} = \max\left\{\frac{1}{\tau_n^{\hat{z}; f_y}} - \xi_n^{z \rightarrow f_y}, \gamma_n^{f_y \rightarrow z}\right\}. \quad (12.36)$$

- If EPC is used, we have

$$\xi_n^{f_y \rightarrow z} = \frac{1}{\tau_n^{\hat{z}; f_y}} - \xi_n^{z \rightarrow f_y}. \quad (12.37)$$

After that, we update the other parameter as

$$\nu_n^{f_y \rightarrow z} = (\xi_n^{f_y \rightarrow z} + \xi_n^{z \rightarrow f_y}) m_n^{\hat{z}; f_y} - \nu_n^{z \rightarrow f_y}. \quad (12.38)$$

12.5.2 $f_{x_k} \rightarrow x_k$

The belief at f_{x_k} is

$$b_k^{f_x}(x_k) \propto f_{x_k}(x_k) \mu_k^{x \rightarrow f_x}(x_k). \quad (12.39)$$

Its mean and variance are computed as

$$\begin{aligned} m_k^{\hat{x}; f_x} &= \mathbb{E}_{b_k^{f_x}}[x_k] \\ \tau_k^{\hat{x}; f_x} &= \text{Var}_{b_k^{f_x}}[x_k]. \end{aligned} \quad (12.40)$$

The threshold $\gamma_k^{f_x \rightarrow x}$ can be obtained by investigating $b_k^{\hat{x}}(x_k)$,

$$\begin{aligned} \gamma_k^{f_x \rightarrow x} + \xi_k^{f_x \rightarrow x} &> 0 \\ \Rightarrow \gamma_k^{f_x \rightarrow x} &= -\xi_k^{f_x \rightarrow x} + \epsilon \end{aligned} \quad (12.41)$$

After that, we can obtain the update for message $\mu_k^{f_x \rightarrow x}$.

12.5.3 $f_z \rightarrow x_k$

We look at the belief at f_z

$$b^{f_z}(z, \mathbf{x}) \propto \delta(z - \mathbf{A}\mathbf{x}) \prod_n \mu_n^{z \rightarrow f_z}(z_n) \prod_k \mu_k^{x \rightarrow f_z}(x_k). \quad (12.42)$$

Marginalize it over $\mathbf{z}$ results to

$$\int b^{f_z}(z, \mathbf{x}) d\mathbf{z} \propto \overline{\mathcal{N}}(\mathbf{x} | \mathbf{m}^{\hat{x}; f_z}, \mathbf{C}^{\hat{x}; f_z}), \quad (12.43)$$

where

$$\begin{aligned} \mathbf{C}^{\hat{x}; f_z} &= [\mathbf{A}^T (\mathbf{C}^{z \rightarrow f_z})^{-1} \mathbf{A} + (\mathbf{C}^{x \rightarrow f_z})^{-1}]^{-1} \\ \mathbf{m}^{\hat{x}; f_z} &= \mathbf{C}^{\hat{x}; f_z} [\mathbf{A}^T (\mathbf{C}^{z \rightarrow f_z})^{-1} \mathbf{m}^{z \rightarrow f_z} + (\mathbf{C}^{x \rightarrow f_z})^{-1} \mathbf{m}^{x \rightarrow f_z}] \end{aligned} \quad (12.44)$$

Denote $\mathbf{e}_k$ as a unit vector whose k -th entry is one, and the marginalized belief is

$$b_k^{\hat{x}; f_z}(x_k) = \int \int b^{f_z}(z, \mathbf{x}) d\mathbf{z} \prod_{k' \neq k} dx_{k'} = \overline{\mathcal{N}}(x_k | m_k^{\hat{x}; f_z}, \tau_k^{\hat{x}; f_z}), \quad (12.45)$$

where

$$\begin{aligned} \tau_k^{\hat{x}; f_z} &= \mathbf{e}_k^T \mathbf{C}^{\hat{x}; f_z} \mathbf{e}_k \\ m_k^{\hat{x}; f_z} &= \mathbf{e}_k^T \mathbf{m}^{\hat{x}; f_z}. \end{aligned} \quad (12.46)$$

To derive the threshold $\gamma_k^{f_z \rightarrow x}$, we look at the belief at x_k :

$$\begin{aligned} \gamma_k^{f_z \rightarrow x} + \xi_k^{f_z \rightarrow x} &> 0 \\ \Rightarrow \gamma_k^{f_z \rightarrow x} &= -\xi_k^{f_z \rightarrow x} + \epsilon. \end{aligned} \quad (12.47)$$

From this point, we can compute the feed back message by either of the methods:

$$\mu_k^{f_z \rightarrow x}(x_k) = \overline{\mathcal{M}}(x_k | \nu_k^{f_z \rightarrow x}, \xi_k^{f_z \rightarrow x}). \quad (12.48)$$

12.5.4 $f_z \rightarrow z_n$

We look at the belief at f_z .

$$b^{f_z}(z, \mathbf{x}) \propto \delta(z - \mathbf{A}\mathbf{x}) \prod_n \mu_n^{z \rightarrow f_z}(z_n) \prod_k \mu_k^{x \rightarrow f_z}(x_k) \quad (12.49)$$

We investigate the marginal belief of z_n first,

$$b_n^{\hat{z}; f_z}(z_n) \propto \int \int b^{f_z}(z, \mathbf{x}) \prod_{n' \neq n} dz_{n'} d\mathbf{x}. \quad (12.50)$$

Use Lemma 6 to integrate $b^{f_z}(z_n, \mathbf{x}) d\mathbf{x}$ on the second line of (12.51):

$$\begin{aligned} b_n^{\hat{z}; f_z}(z_n) &\propto \overline{\mathcal{N}}(\mathbf{m}_{\bar{n}}^{z \rightarrow f_z} | \mathbf{A}_{\bar{n}} \mathbf{m}^{x \rightarrow f_z}, \mathbf{C}_{\bar{n}}^{z \rightarrow f_z} + \mathbf{A}_{\bar{n}} \mathbf{C}^{x \rightarrow f_z} \mathbf{A}_{\bar{n}}^T) \\ &\cdot \int \delta(z_n - \mathbf{a}_n^T \mathbf{x}) \overline{\mathcal{N}}(z_n | m_n^{z \rightarrow f_z}, \tau_n^{z \rightarrow f_z}) \overline{\mathcal{N}}(\mathbf{x} | \mathbf{m}_n^{\tilde{x}}, \mathbf{C}_n^{\tilde{x}}) d\mathbf{x} \\ &\propto \overline{\mathcal{N}}(z_n | \mathbf{a}_n^T \mathbf{m}^{\hat{x}; f_z}, \mathbf{a}_n^T \mathbf{C}^{\hat{x}; f_z} \mathbf{a}_n) \end{aligned} \quad (12.51)$$

where

$$\begin{aligned} \mathbf{C}_n^{\tilde{x}} &= [\mathbf{A}_{\bar{n}}^T (\mathbf{C}_{\bar{n}}^{z \rightarrow f_z})^{-1} \mathbf{A}_{\bar{n}} + (\mathbf{C}^{x \rightarrow f_z})^{-1}]^{-1}. \\ \mathbf{m}_n^{\tilde{x}} &= \mathbf{C}_n^{\tilde{x}} [\mathbf{A}_{\bar{n}}^T (\mathbf{C}_{\bar{n}}^{z \rightarrow f_z})^{-1} \mathbf{m}_{\bar{n}}^{z \rightarrow f_z} + (\mathbf{C}^{x \rightarrow f_z})^{-1} \mathbf{m}^{x \rightarrow f_z}] \end{aligned} \quad (12.52)$$

We also have

$$\mathbf{a}_n^T \mathbf{C}^{\hat{x}; f_z} \mathbf{a}_n = \mathbf{e}_n^T [(\mathbf{A} \mathbf{C}^{x \rightarrow f_z} \mathbf{A}^T)^{-1} + (\mathbf{C}^{z \rightarrow f_z})^{-1}]^{-1} \mathbf{e}_n \quad (12.53)$$

if the inversion exists.

An interesting relation is $\forall \boldsymbol{\nu}$:

$$\mathbf{a}_n^T \mathbf{C}_n^{\tilde{x}} \boldsymbol{\nu} = \tau_n^{z \rightarrow f_z} (\mathbf{a}_n^T \mathbf{C}^{\hat{x}; f_z} \mathbf{a}_n - \tau_n^{z \rightarrow f_z})^{-1} \mathbf{a}_n^T \mathbf{C}^{\hat{x}; f_z} \boldsymbol{\nu} \quad (12.54)$$

In fact, $\overline{\mathcal{N}}(\mathbf{x} | \mathbf{m}_n^{\tilde{x}}, \mathbf{C}_n^{\tilde{x}})$ is an approximation of $p(\mathbf{x} | \forall n' \neq n : y_{n'})$.

The mean and variance of the marginal belief of z_n at f_z is

$$\begin{aligned} m_n^{\hat{z}; f_z} &= \mathbb{E}_{b_n^{\hat{z}; f_z}}[z_n] = \mathbf{a}_n^T \mathbf{m}^{\hat{x}; f_z} \\ \tau_n^{\hat{z}; f_z} &= \text{Var}_{b_n^{\hat{z}; f_z}}[z_n] = \mathbf{a}_n^T \mathbf{C}^{\hat{x}; f_z} \mathbf{a}_n \end{aligned} \quad (12.55)$$

To derive the threshold, we observe the destination node of message $\mu_n^{f_z \rightarrow z}$. In order to make the belief at $b_n^{\hat{z}}(z_n)$ integrable, we have

$$\begin{aligned} \xi_n^{f_y \rightarrow z} + \gamma_n^{f_z \rightarrow z} &> 0 \\ \gamma_n^{f_z \rightarrow z} &= -\xi_n^{f_y \rightarrow z} + \epsilon \end{aligned} \quad (12.56)$$

Based on (12.55) and (12.56), we can update the message $\mu_n^{f_z \rightarrow z}(z_n)$ based on the EP algorithm discussed before.

12.5.5 $z \rightarrow f_y$

The belief at z_n is

$$b_n^{\hat{z}}(z_n) = \mathcal{N}(z_n | m_n^{\hat{z}}, \tau_n^{\hat{z}}) \propto \mu_n^{f_z \rightarrow z}(z_n) \mu_n^{f_y \rightarrow z}(z_n), \quad (12.57)$$

where

$$\begin{aligned} \tau_n^{\hat{z}} &= (\xi_n^{f_z \rightarrow z} + \xi_n^{f_y \rightarrow z})^{-1} \\ m_n^{\hat{z}} &= \tau_n^{\hat{z}} (\nu_n^{f_z \rightarrow z} + \nu_n^{f_y \rightarrow z}). \end{aligned} \quad (12.58)$$

We then determine the threshold precision $\gamma_n^{z \rightarrow f_y}$ based on the factor $f_{y_n}(z_n)$. For example, if $f_{y_n}(z_n)$ is proportional to a Gaussian Mixture Model, the sum of $\gamma_n^{z \rightarrow f_y}$ and the precision of the smallest GMM elements should be greater than zero.

12.5.6 $z \rightarrow f_z$

We need to determine the threshold precision $\gamma_n^{z \rightarrow f_z}$. The update of one message correspond to a rank-one update of the covariance matrix $\mathbf{C}^{\hat{\mathbf{x}}; f_z}$

$$(\mathbf{C}^{\hat{\mathbf{x}}; f_z})^{-1} + \mathbf{a}_n(\gamma_n^{z \rightarrow f_z} - \xi_n^{z \rightarrow f_z})\mathbf{a}_n > 0, \quad (12.59)$$

where $\xi_n^{z \rightarrow f_z}$ is the message variance before updating and $\mathbf{C}^{\hat{\mathbf{x}}; f_z}$ is the covariance matrix before updating.

From Lemma 7, the positive definite condition is equivalent to

$$\gamma_n^{z \rightarrow f_z} > \xi_n^{z \rightarrow f_z} - (\mathbf{a}_n^T \mathbf{C}^{\hat{\mathbf{x}}; f_z} \mathbf{a}_n)^{-1} = \xi_n^{z \rightarrow f_z} - \xi_n^{\hat{z}; f_z}. \quad (12.60)$$

Proposition 3. *In ACEP, with proper initialization (e.g. all positive message precision), if the update of $\mu_n^{f_z \rightarrow z}$, $\mu_n^{f_y \rightarrow z}$, $\mu_n^{z \rightarrow f_y}$ and $\mu_n^{z \rightarrow f_z}$ follows the following order:*

$$\mu_n^{f_z \rightarrow z} \rightarrow \mu_n^{z \rightarrow f_y} \rightarrow \mu_n^{f_y \rightarrow z} \rightarrow \mu_n^{z \rightarrow f_z}, \quad (12.61)$$

the update of $\mu_n^{f_z \rightarrow z}$ and $\mu_n^{z \rightarrow f_z}$ does not need to evaluate the threshold.

Proof. We can prove this by mathematical induction. We use superscript t to denote the iteration number. Assume the proposition hold till $t - 1$ iteration of (12.61). By the induction hypothesis, $(\mu_n^{f_y \rightarrow z})^{(t-1)} = (\mu_n^{z \rightarrow f_z})^{(t-1)}$. Therefore, $\xi_n^{f_z \rightarrow z}$ is updated by

$$\begin{aligned} (\xi_n^{f_z \rightarrow z})^{(t)} + (\xi_n^{z \rightarrow f_z})^{(t-1)} &= \frac{1}{\mathbf{a}_n^T (\mathbf{C}^{\hat{\mathbf{x}}; f_z})^{(t-1)} \mathbf{a}_n} \\ &= (\xi_n^{f_z \rightarrow z})^{(t)} + (\mu_n^{f_y \rightarrow z})^{(t-1)} > 0, \end{aligned} \quad (12.62)$$

where we denote the covariance matrix at b^{f_z} updated just before (12.61) as $(\mathbf{C}^{\hat{\mathbf{x}}; f_z})^{(t-1)}$, and the one after the chain (12.61) as $(\mathbf{C}^{\hat{\mathbf{x}}; f_z})^{(t)}$. The second line of (12.62) is a result of induction hypothesis and Lemma 7.

Since (12.62) is greater than zero, $(\xi_n^{f_z \rightarrow z})^{(t)}$ meets the threshold automatically.

Based on the update rule for $\mu_n^{f_y \rightarrow z}$,

$$(\xi_n^{f_y \rightarrow z})^{(t)} + (\xi_n^{f_z \rightarrow z})^{(t)} > 0. \quad (12.63)$$

Now we investigate the constraint for $(\xi_n^{z \rightarrow fz})^{(t)}$. We look at $(\mathbf{C}^{\hat{\mathbf{x}}; fz})^{(t)}$ just after the update chain (12.61):

$$\begin{aligned} & (\mathbf{C}^{\hat{\mathbf{x}}; fz})^{(t)} \\ &= [(\mathbf{C}^{\hat{\mathbf{x}}; fz})^{(t-1)}]^{-1} + \mathbf{a}_n [(\xi_n^{z \rightarrow fz})^{(t)} - (\xi_n^{z \rightarrow fz})^{(t-1)}] \mathbf{a}_n^T]^{-1}. \end{aligned} \quad (12.64)$$

The covariance matrix (12.64) need to remain positive definite. By Lemma 7, positive definiteness of (12.64) is equivalent to the following constraint for $(\xi_n^{z \rightarrow fz})^{(t)}$

$$\begin{aligned} (\xi_n^{z \rightarrow fz})^{(t)} &> -\frac{1}{\mathbf{a}_n^T (\mathbf{C}^{\hat{\mathbf{x}}; fz})^{(t-1)} \mathbf{a}_n} + (\xi_n^{z \rightarrow fz})^{(t-1)} \\ &\Leftrightarrow (\xi_n^{z \rightarrow fz})^{(t)} + (\xi_n^{fz \rightarrow z})^{(t)} > 0, \end{aligned} \quad (12.65)$$

where the second line follows from (12.62). Compare the update of $\mu_n^{z \rightarrow fz}$, (12.63) with condition (12.65), and we find out that in ACEP, we can directly set $\mu_n^{z \rightarrow fz} = \mu_n^{fz \rightarrow z}$ without checking the threshold. $\square$

Remark. The previous proposition indicates that in EPC, the belief b^{fz} is always integrable.

Therefore, if we have the update order (12.61), the update for $\mu_n^{z \rightarrow fz}$ in both ACEP and EPC is:

$$\mu_n^{z \rightarrow fz}(z_n) = \mu_n^{fz \rightarrow z}(z_n) \quad (12.66)$$

Otherwise, if we have different order, we apply the constrained KLD optimization in Section 12.2.1 based on (12.58) and (12.60).

12.5.7 $x \rightarrow f_x$

The belief at x_k is

$$b_k^{\hat{x}}(x_k) \propto \mu_k^{f_x \rightarrow x}(x_k) \mu_k^{f_z \rightarrow x}. \quad (12.67)$$

The belief is always integrable. The mean and variance of $b_k^{\hat{x}}$ are

$$\begin{aligned} \tau_k^{\hat{x}} &= (\xi_k^{f_x \rightarrow x} + \xi_k^{f_z \rightarrow x})^{-1} \\ m_k^{\hat{x}} &= \tau_k^{\hat{x}} (\nu_k^{f_x \rightarrow x} + \nu_k^{f_z \rightarrow x}). \end{aligned} \quad (12.68)$$

To derive the threshold $\gamma_k^{x \rightarrow fz}$, we observe the belief at f_{x_k} . If GMM prior is used, the sum of $\gamma_k^{x \rightarrow fz}$ and the minimum component precision should be greater than zero.

12.5.8 $x \rightarrow f_z$

To determine the threshold $\gamma_k^{x \rightarrow fz}$, we observe the rank one update. If the precision of message $\mu_k^{x \rightarrow fz}$ is updated from $\xi_k^{x \rightarrow fz}$ to $\gamma_k^{x \rightarrow fz}$, the new covariance matrix of belief b^{fz} is updated by

$$[(\mathbf{C}^{\hat{\mathbf{x}}; fz})^{-1} + \mathbf{e}_k (\gamma_k^{x \rightarrow fz} - \xi_k^{x \rightarrow fz}) \mathbf{e}_k^T]^{-1}, \quad (12.69)$$

where $\mathbf{e}_k$ denotes a unit vector with the k -th entry equal to one. Based on relation (12.46) and Lemma 7, (12.69) is positive definite iff.

$$\gamma_k^{x \rightarrow fz} > \xi_k^{x \rightarrow fz} - \xi_k^{\hat{x}; fz}. \quad (12.70)$$

Proposition 4. In ACEP, with proper initialization (e.g. all positive message precision), if the update of $\mu_k^{fz \rightarrow x}$, $\mu_k^{fx \rightarrow x}$, $\mu_k^{x \rightarrow fx}$ and $\mu_k^{x \rightarrow fz}$ follows the following order:

$$\mu_k^{fz \rightarrow x} \rightarrow \mu_k^{x \rightarrow fx} \rightarrow \mu_k^{fx \rightarrow x} \rightarrow \mu_k^{x \rightarrow fz}, \quad (12.71)$$

the update of $\mu_k^{fz \rightarrow x}$ and $\mu_k^{x \rightarrow fz}$ does not need to evaluate the threshold.

Proof. The proof is analog to the one of Proposition 3. $\square$

Therefore, if we have the update order (12.71), the update

$$\mu_k^{x \rightarrow fz}(x_k) = \mu_k^{fx \rightarrow x}(x_k) \quad (12.72)$$

automatically guaranties that b^{fz} is integrable.

12.6 Useful Relations

An important relation based on Gaussian reproduction lemma is

$$\begin{aligned} & \bar{\mathcal{N}}(\mathbf{H}\mathbf{x}|\mathbf{a}, \mathbf{A})\bar{\mathcal{N}}(\mathbf{x}|\mathbf{b}, \mathbf{B}) \\ &= \bar{\mathcal{N}}(\mathbf{x}|\mathbf{c}, \mathbf{C})\bar{\mathcal{N}}(\mathbf{a}|\mathbf{H}\mathbf{b}, \mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{A}), \end{aligned} \quad (12.73)$$

where

$$\begin{aligned} \mathbf{C} &= (\mathbf{H}^T \mathbf{A}^{-1} \mathbf{H} + \mathbf{B}^{-1})^{-1} \\ \mathbf{c} &= \mathbf{C}(\mathbf{H}^T \mathbf{A}^{-1} \mathbf{a} + \mathbf{B}^{-1} \mathbf{b}). \end{aligned} \quad (12.74)$$

Lemma 6. Let $\mathbf{a} \in \mathbb{R}^K$, $\mathbf{m}_x \in \mathbb{R}^K$, and $\mathbf{C}_x \in \mathbb{R}^{K \times K}$ be symmetric. If

$$\int \int \bar{\mathcal{N}}(\mathbf{x}|\mathbf{m}_x, \mathbf{C}_x) \bar{\mathcal{N}}(z|m_z, \tau_z) \delta(z - \mathbf{a}^T \mathbf{x}) dz d\mathbf{x} \quad (12.75)$$

is integrable, then

$$\begin{aligned} & \int \bar{\mathcal{N}}(\mathbf{x}|\mathbf{m}_x, \mathbf{C}_x) \bar{\mathcal{N}}(z|m_z, \tau_z) \delta(z - \mathbf{a}^T \mathbf{x}) d\mathbf{x} \\ &= \bar{\mathcal{N}}(z|m_{\hat{z}}, \tau_{\hat{z}}) \sqrt{(2\pi)^{K-1} (\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} \det(\mathbf{C}_x)} \\ & \quad \cdot \bar{\mathcal{N}}(0|m_z - \mathbf{a}^T \mathbf{m}_x, \mathbf{a}^T \mathbf{C}_x \mathbf{a} + \tau_z), \end{aligned} \quad (12.76)$$

where

$$\begin{aligned} \tau_{\hat{z}} &= [(\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} + \tau_z^{-1}]^{-1} \\ m_{\hat{z}} &= \tau_{\hat{z}} [(\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} \mathbf{a}^T \mathbf{m}_x + \tau_z^{-1} m_z] \end{aligned} \quad (12.77)$$

Proof. We compute the integration by substitution. Define

$$\mathbf{U} = [\mathbf{u}_1 \quad \dots \quad \mathbf{u}_{K-1}] \in \mathbb{R}^{K \times (K-1)}, \quad (12.78)$$

such that $\forall k, k' \in [1, K-1] : \mathbf{a}^T \mathbf{u}_k = 0$ and $\mathbf{u}_k^T \mathbf{u}_{k'} = \delta_{kk'}$, where $\delta_{kk'}$ denotes Kronecker delta that evaluates to zero if $k' \neq k$ and evaluates to one if $k = k'$.

We construct a full rank basis matrix

$$\mathbf{H} = [\mathbf{a} \quad \mathbf{U}] \in \mathbb{R}^{K \times K}. \quad (12.79)$$

We substitute $\mathbf{x}$ by $\boldsymbol{\theta}$, where $\mathbf{x} = \mathbf{H}\boldsymbol{\theta}$ and thus, $d\mathbf{x} = |\det(\mathbf{H})|d\boldsymbol{\theta}$.
The integral in (12.76) can be computed as

$$\begin{aligned}
& \int \mathcal{N}(\mathbf{x}|\mathbf{m}_x, \mathbf{C}_x) \mathcal{N}(z|m_z, \tau_z) \delta(z - \mathbf{a}^T \mathbf{x}) d\mathbf{x} \\
&= \int \int \mathcal{N}(\mathbf{a}\theta_1 + \mathbf{U}\bar{\boldsymbol{\theta}}|\mathbf{m}_x, \mathbf{C}_x) d\bar{\boldsymbol{\theta}} \\
&\quad \cdot \mathcal{N}(z|m_z, \tau_z) \delta(z - |\mathbf{a}|^2 \theta_1) |\det(\mathbf{H})| d\theta_1 \\
&= \int \int \mathcal{N}(\bar{\boldsymbol{\theta}}|\mathbf{m}_{\bar{\boldsymbol{\theta}}}(\theta_1), \mathbf{C}_{\bar{\boldsymbol{\theta}}}) d\bar{\boldsymbol{\theta}} \exp\left[\frac{1}{2} \mathbf{m}_{\bar{\boldsymbol{\theta}}}^T(\theta_1) \mathbf{C}_{\bar{\boldsymbol{\theta}}}^{-1} \mathbf{m}_{\bar{\boldsymbol{\theta}}}(\theta_1)\right] \\
&\quad \cdot \exp\left[-\frac{1}{2} (\mathbf{m}_x - \mathbf{a}\theta_1)^T \mathbf{C}_x^{-1} (\mathbf{m}_x - \mathbf{a}\theta_1)\right] \\
&\quad \cdot \mathcal{N}(z|m_z, \tau_z) \delta(z - |\mathbf{a}|^2 \theta_1) \cdot |\det(\mathbf{H})| d\theta_1 \\
&= \sqrt{(2\pi)^{K-1} \det(\mathbf{C}_{\bar{\boldsymbol{\theta}}})} \int \exp\left[\frac{1}{2} \mathbf{m}_{\bar{\boldsymbol{\theta}}}^T(\theta_1) \mathbf{C}_{\bar{\boldsymbol{\theta}}}^{-1} \mathbf{m}_{\bar{\boldsymbol{\theta}}}(\theta_1)\right] \\
&\quad \cdot \exp\left[-\frac{1}{2} (\mathbf{m}_x - \mathbf{a}\theta_1)^T \mathbf{C}_x^{-1} (\mathbf{m}_x - \mathbf{a}\theta_1)\right] \\
&\quad \cdot \mathcal{N}(z|m_z, \tau_z) \delta(z - |\mathbf{a}|^2 \theta_1) \cdot |\det(\mathbf{H})| d\theta_1
\end{aligned} \tag{12.80}$$

where

$$\begin{aligned}
\mathbf{C}_{\bar{\boldsymbol{\theta}}} &= (\mathbf{U}^T \mathbf{C}_x^{-1} \mathbf{U})^{-1} \\
\mathbf{m}_{\bar{\boldsymbol{\theta}}}(\theta_1) &= \mathbf{C}_{\bar{\boldsymbol{\theta}}} [\mathbf{U}^T \mathbf{C}_x^{-1} (\mathbf{m}_x - \mathbf{a}\theta_1)]
\end{aligned} \tag{12.81}$$

The integral of (12.75) exists, which implies that we have positive definite matrix

$$\mathbf{C}_{\bar{\boldsymbol{\theta}}} > 0. \tag{12.82}$$

From this point, we only have scalar variables θ_1 and z .

Continuing from (12.80), we have

$$\begin{aligned}
& \int \mathcal{N}(\mathbf{x}|\mathbf{m}_x, \mathbf{C}_x) \mathcal{N}(z|m_z, \tau_z) \delta(z - \mathbf{a}^T \mathbf{x}) d\mathbf{x} \\
&= \frac{|\det(\mathbf{H})| \sqrt{(2\pi)^{K-1} \det(\mathbf{C}_{\bar{\boldsymbol{\theta}}})}}{|\mathbf{a}|^2}
\end{aligned} \tag{12.83}$$

$$\cdot \exp\left[-\frac{1}{2} z^T (\mathbf{b}_3^T \mathbf{C}_x^{-1} \mathbf{b}_3 - \mathbf{b}_1^T \mathbf{C}_{\bar{\boldsymbol{\theta}}}^{-1} \mathbf{b}_1 + \tau_z^{-1}) z\right] \tag{12.84}$$

$$\cdot \exp\left[z^T (\mathbf{b}_3^T \mathbf{C}_x^{-1} \mathbf{m}_x - \mathbf{b}_1^T \mathbf{C}_{\bar{\boldsymbol{\theta}}}^{-1} \mathbf{b}_2 + \tau_z^{-1} m_z)\right] \tag{12.85}$$

$$\cdot \exp\left[-\frac{1}{2} (\mathbf{m}_x^T \mathbf{C}_x^{-1} \mathbf{m}_x - \mathbf{b}_2^T \mathbf{C}_{\bar{\boldsymbol{\theta}}}^{-1} \mathbf{b}_2 + m_z^T \tau_z^{-1} m_z)\right], \tag{12.86}$$

where

$$\begin{aligned}
\mathbf{b}_1 &= \mathbf{C}_{\bar{\boldsymbol{\theta}}} \mathbf{U}^T \mathbf{C}_x^{-1} \frac{\mathbf{a}}{|\mathbf{a}|^2} \\
\mathbf{b}_2 &= \mathbf{C}_{\bar{\boldsymbol{\theta}}} \mathbf{U}^T \mathbf{C}_x^{-1} \mathbf{m}_x \\
\mathbf{b}_3 &= \frac{\mathbf{a}}{|\mathbf{a}|^2};
\end{aligned} \tag{12.87}$$

We try to simplify (12.84). Let

$$\mathbf{Q} = \begin{bmatrix} \mathbf{a}^T/|\mathbf{a}| \\ \mathbf{U}^T \end{bmatrix} \mathbf{C}_x^{-1} \begin{bmatrix} \mathbf{a}/|\mathbf{a}| & \mathbf{U} \end{bmatrix} \quad (12.88)$$

be a matrix of 2 blocks by 2 blocks. Since $\begin{bmatrix} \mathbf{a}/|\mathbf{a}| & \mathbf{U} \end{bmatrix}$ is a unitary matrix, we have

$$\mathbf{Q}^{-1} = \begin{bmatrix} \mathbf{a}^T/|\mathbf{a}| \\ \mathbf{U}^T \end{bmatrix} \mathbf{C}_x \begin{bmatrix} \mathbf{a}/|\mathbf{a}| & \mathbf{U} \end{bmatrix} \quad (12.89)$$

According to Schur complement and matrix inversion lemma, the first two terms of the quadratic coefficient in (12.84) can be verified to be

$$\mathbf{b}_3^T \mathbf{C}_x^{-1} \mathbf{b}_3 - \mathbf{b}_1^T \mathbf{C}_{\bar{\theta}}^{-1} \mathbf{b}_1 = \frac{1}{|\mathbf{a}|^2} ([\mathbf{Q}^{-1}]_{11})^{-1} = (\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} \quad (12.90)$$

Now we try to simplify (12.85). The first two constant terms can be expanded as

$$\mathbf{b}_3^T \mathbf{C}_x^{-1} \mathbf{m}_x - \mathbf{b}_1^T \mathbf{C}_{\bar{\theta}}^{-1} \mathbf{b}_2 = \mathbf{b}_4^T \mathbf{m}_x. \quad (12.91)$$

where

$$\mathbf{b}_4^T = \frac{\mathbf{a}^T}{|\mathbf{a}|^2} \mathbf{C}_x^{-1} - \frac{\mathbf{a}^T}{|\mathbf{a}|^2} \mathbf{C}_x^{-1} \mathbf{U} (\mathbf{U}^T \mathbf{C}_x^{-1} \mathbf{U})^{-1} \mathbf{U}^T \mathbf{C}_x^{-1}. \quad (12.92)$$

It has the following two properties

- $\mathbf{b}_4^T \frac{\mathbf{a}}{|\mathbf{a}|^2}$ is the same as expanding the left hand side of (12.90). Therefore,

$$\mathbf{b}_4^T \frac{\mathbf{a}}{|\mathbf{a}|^2} = (\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1}. \quad (12.93)$$

- It is in the null space of $\mathbf{U}$:

$$\mathbf{b}_4^T \mathbf{U} = 0 \quad (12.94)$$

Therefore, from (12.94), we know $\mathbf{b}_4$ must be a scalar multiple of $\mathbf{a}$. By (12.93), we have

$$\mathbf{b}_4^T = (\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} \mathbf{a}^T. \quad (12.95)$$

We investigate (12.86), the first two terms can be expanded as

$$\mathbf{m}_x^T \mathbf{C}_x^{-1} \mathbf{m}_x - \mathbf{b}_2^T \mathbf{C}_{\bar{\theta}}^{-1} \mathbf{b}_2 = \mathbf{b}_5^T \mathbf{m}_x, \quad (12.96)$$

where

$$\mathbf{b}_5^T = \mathbf{m}_x^T \mathbf{C}_x^{-1} - \mathbf{m}_x^T \mathbf{C}_x^{-1} \mathbf{U} (\mathbf{U}^T \mathbf{C}_x^{-1} \mathbf{U})^{-1} \mathbf{U}^T \mathbf{C}_x^{-1} \quad (12.97)$$

Similar as before, $\mathbf{b}_5^T$ has the following two properties

- Compare (12.92) with (12.97), we can verify that

$$\mathbf{b}_5^T \frac{\mathbf{a}}{|\mathbf{a}|^2} = (\mathbf{b}_4^T \mathbf{m}_x)^T = \mathbf{m}_x^T \mathbf{a} (\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} \quad (12.98)$$

- It is also in the null space of $\mathbf{U}$

$$\mathbf{b}_5^T \mathbf{U} = 0 \quad (12.99)$$

Therefore, $\mathbf{b}_5^T$ must be a scalar multiple of $\mathbf{a}$. From (12.98), we have

$$\mathbf{b}_5^T = \mathbf{m}_x^T \mathbf{a} (\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} \mathbf{a}^T \quad (12.100)$$

Finally, we examine the determinant in (12.83). Since

$$\mathbf{H} = \begin{bmatrix} \mathbf{a}/|\mathbf{a}| & \mathbf{U} \end{bmatrix} \begin{bmatrix} |\mathbf{a}| & \\ & \mathbf{I} \end{bmatrix}, \quad (12.101)$$

we have $|\det(\mathbf{H})| = |\mathbf{a}|$.

From (12.88), we have

$$\mathbf{C}_{\bar{\theta}} = ([\mathbf{Q}]_{22})^{-1}, \quad (12.102)$$

where $[\mathbf{Q}]_{22}$ denote the matrix at the second block row and the second block column of $\mathbf{Q}$. Based on block determinant property and (12.90), we have

$$\det(\mathbf{C}_{\bar{\theta}}) = \det(\mathbf{C}_x) (\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} |\mathbf{a}|^2 \quad (12.103)$$

Substitute the discussion results into (12.83) - (12.86) and we have

$$\begin{aligned} & \int \mathcal{N}(\mathbf{x} | \mathbf{m}_x, \mathbf{C}_x) \mathcal{N}(z | m_z, \tau_z) \delta(z - \mathbf{a}^T \mathbf{x}) d\mathbf{x} \\ &= \sqrt{(2\pi)^{K-1} (\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} \det(\mathbf{C}_x)} \\ & \quad \cdot \mathcal{N}(0 | m_z - \mathbf{a}^T \mathbf{m}_x, \mathbf{a}^T \mathbf{C}_x \mathbf{a} + \tau_z) \\ & \quad \cdot \mathcal{N}(z | m_{\hat{z}}, \tau_{\hat{z}}), \end{aligned} \quad (12.104)$$

where

$$\begin{aligned} \tau_{\hat{z}} &= [(\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} + \tau_z^{-1}]^{-1} \\ m_{\hat{z}} &= \tau_{\hat{z}} [(\mathbf{a}^T \mathbf{C}_x \mathbf{a})^{-1} \mathbf{a}^T \mathbf{m}_x + \tau_z^{-1} m_z] \end{aligned} \quad (12.105)$$

□

Lemma 7. Let $\mathbf{C}^{-1} \in \mathbb{R}^{K \times K}$ be a symmetric positive definite matrix, then $(\mathbf{C}^{-1} + \mathbf{a} \tau^{-1} \mathbf{a}^T)^{-1}$ is also positive definite iff. $\tau^{-1} > -\frac{1}{\mathbf{a}^T \mathbf{C} \mathbf{a}}$.

Proof. The matrix inversion does not change the positive definiteness. Therefore, we investigate $\mathbf{C}^{-1} + \mathbf{a} \tau^{-1} \mathbf{a}^T$. Since $\mathbf{C}^{-1}$ is positive definite and symmetric, we factorize it by

$$\mathbf{C}^{-1} = \mathbf{C}^{-1/2} \mathbf{C}^{-T/2}, \quad (12.106)$$

such that $\mathbf{C}^{-1/2}$ is also a symmetric positive definite matrix. Therefore, we have

$$\mathbf{C}^{-1} + \mathbf{a} \tau^{-1} \mathbf{a}^T = \mathbf{C}^{-1/2} (\mathbf{I} + \mathbf{b} \tau^{-1} \mathbf{b}^T) \mathbf{C}^{-T/2} \quad (12.107)$$

where

$$\mathbf{b} = \mathbf{C}^{1/2} \mathbf{a}. \quad (12.108)$$

Since $\mathbf{C}^{-1/2}$ is full-rank, (12.107) is positive definite iff. $(\mathbf{I} + \mathbf{b} \tau^{-1} \mathbf{b}^T)$ is positive definite. We construct a unitary matrix $\mathbf{U} = [\mathbf{u}_1 \ \dots \ \mathbf{u}_{K-1}] \in \mathbb{R}^{K \times (K-1)}$, such that $\forall k, k' \in [1, K-1] : \mathbf{b}^T \mathbf{u}_k = 0$ and $\mathbf{u}_k^T \mathbf{u}_{k'} = \delta_{kk'}$, where $\delta_{kk'}$ denotes Kronecker delta that evaluates to zero if $k' \neq k$ and evaluates to one if $k = k'$.

Another square unitary matrix can be constructed by

$$\mathbf{H} = \begin{bmatrix} \frac{\mathbf{b}}{|\mathbf{b}|} & \mathbf{U} \end{bmatrix} \in \mathbb{R}^{K \times K}. \quad (12.109)$$

Therefore,

$$\mathbf{I} + \mathbf{b}\tau^{-1}\mathbf{b}^T = \mathbf{H} \begin{bmatrix} 1 + |\mathbf{b}|^2\tau^{-1} & \\ & \mathbf{I} \end{bmatrix} \mathbf{H}^T. \quad (12.110)$$

From (12.110), $(\mathbf{C}^{-1} + \mathbf{a}\tau^{-1}\mathbf{a}^T)^{-1}$ is positive definite iff.

$$1 + |\mathbf{b}|^2\tau^{-1} > 0 \Leftrightarrow \tau^{-1} > -\frac{1}{\mathbf{a}^T\mathbf{C}\mathbf{a}} \quad (12.111)$$

□

Remark. Lemma 7 implies that it is easier to use precision instead of variance when deriving. When using the precision representation, the feasible set is $\tau^{-1} \in (-\frac{1}{\mathbf{a}^T\mathbf{C}\mathbf{a}}, +\infty)$. However, in variance representation, this condition is equivalent to $\tau \in (-\infty, \mathbf{a}^T\mathbf{C}\mathbf{a}) \cup (0, +\infty)$.

12.7 Gaussian Mixture Model (GMM)

For simulations, we use GMM prior $p(x_k)$ and likelihood $p(y_n|z_n)$. Consider the following belief

$$b(\theta) \propto f(\theta)\mu(\theta), \quad (12.112)$$

where

$$f(\theta) = \sum_s w_s \overline{\mathcal{N}}(\theta|m_s, \tau_s) = \sum_s w_s \overline{\mathcal{M}}(\theta|\nu_s, \xi_s) \quad (12.113)$$

and

$$\mu(\theta) = \overline{\mathcal{N}}(\theta|m, \tau) = \overline{\mathcal{M}}(\theta|\nu, \xi). \quad (12.114)$$

Therefore,

$$b(\theta) = \sum_s w_s \overline{\mathcal{N}}(0|m_s - m, \tau_s + \tau) \overline{\mathcal{M}}(\theta|\nu_s + \nu, \xi_s + \xi). \quad (12.115)$$

Since $b(\theta)$ is integrable, we have $\forall s : \xi_s + \xi > 0$. We can transform the factor containing θ to normalized Gaussian

$$\overline{\mathcal{M}}(\theta|\nu_s + \nu, \xi_s + \xi) = \sqrt{2\pi\tau_{\hat{\theta}|s}} \mathcal{N}(\theta|m_{\hat{\theta}|s}, \tau_{\hat{\theta}|s}), \quad (12.116)$$

where

$$\begin{aligned} \tau_{\hat{\theta}|s} &= \frac{1}{\xi_s + \xi} \\ m_{\hat{\theta}|s} &= \frac{\nu_s + \nu}{\xi_s + \xi}. \end{aligned} \quad (12.117)$$

We rewrite the belief as

$$b(\theta) \propto \sum_s \eta_s \mathcal{N}(\theta|m_{\hat{\theta}|s}, \tau_{\hat{\theta}|s}), \quad (12.118)$$

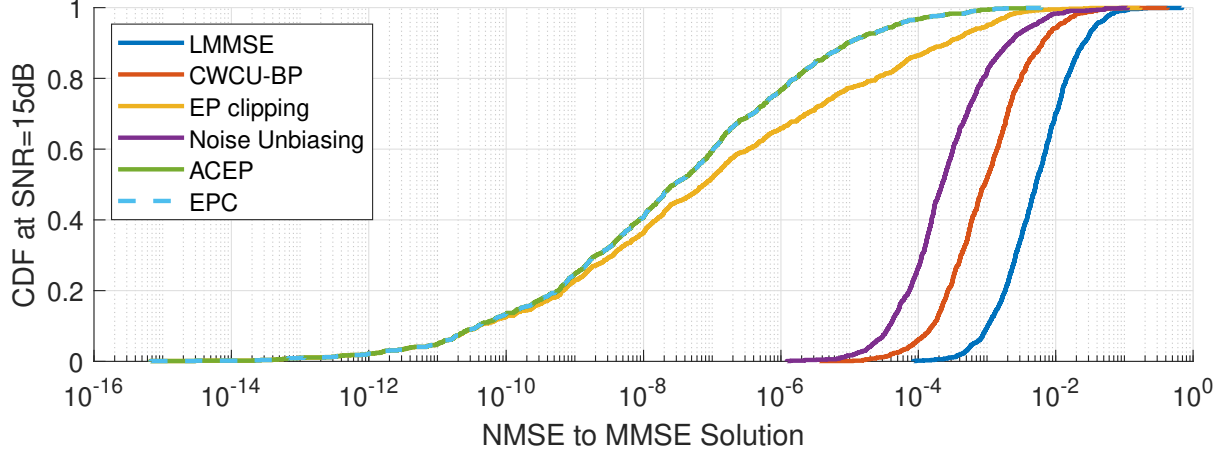


Figure 12.1: CDF of NMSE

where

$$\eta_s = w_s \bar{\mathcal{N}}(0|m_s - m, \tau_s + \tau) \sqrt{2\pi\tau_{\hat{\theta}|s}}. \quad (12.119)$$

We can normalize the belief. Define

$$\rho_s = \frac{\eta_s}{\sum_{s'} \eta_{s'}}, \quad (12.120)$$

and we have

$$b(\theta) = \sum_s \rho_s \mathcal{N}(\theta|m_{\hat{\theta}|s}, \tau_{\hat{\theta}|s}). \quad (12.121)$$

The belief mean and variance are

$$\begin{aligned} m_{\hat{\theta}} &= \mathbb{E}_b[\theta] = \sum_s \rho_s m_{\hat{\theta}|s} \\ \tau_{\hat{\theta}} &= \text{Var}_b[\theta] = \left[\sum_s \rho_s (\tau_{\hat{\theta}|s} + |m_{\hat{\theta}|s}|^2) \right] - |m_{\hat{\theta}}|^2. \end{aligned} \quad (12.122)$$

12.8 Simulation Results

We consider a system of size $N \times K = 8 \times 12$. The prior distribution of each symbol is independent and follows GMM with exponential decay:

$$\begin{aligned} p(x_k) &= \frac{1}{2} \mathcal{N}(x_k | 2^{-n+1}, 0.1 \times 2^{-2(n-1)}) \\ &\quad + \frac{1}{2} \mathcal{N}(x_k | -2^{-n+1}, 0.1 \times 2^{-2(n-1)}). \end{aligned} \quad (12.123)$$

Each entry in the measurement matrix $\mathbf{A}$ is generated from i.i.d. Gaussian. Additive white Gaussian noise is assumed, whose power is adjusted such that the SNR is 15 dB. We generate 1000 independent realizations.

Besides our proposed methods, we also simulated the results for LMMSE, CWCUBP [24, 25], EP clipping [10], and Noise Unbiasing [28]. The simulation results show that the performance of our proposed methods surpasses the existing attempts to handle the non-integrable beliefs. This indicates that in order to have more accurate estimates, we have to accept the non-integrable messages.

12.9 Conclusions

We proposed two EP frameworks that avoid non-integrable beliefs without shrinking the feasible set. Both methods permits non-integrable messages and factors. The proposed methods are then applied to GLM. Based on our analysis, we find out that if EPC is used, the only beliefs that require checking are $b_n^{f_y}$ and $b_k^{f_x}$. On the other hand, if ACEP is used, only the messages to and from factor nodes f_{y_n} and f_{x_k} need to be compared with the thresholds to guarantee integrable beliefs. The simulation results indicate that in order to achieve better results, we must accept non-integrable messages.

Chapter 13

Conclusions

In this thesis, we investigate message-passing-based algorithms for both linear and bilinear estimation. In the first part, we apply EP to a loop-free factorized generalized linear model (GLM). This enables the construction of equivalent Gaussian priors and observation noise that preserve the posterior means and variances of the input signals. The resulting extrinsic messages can be shown to correspond to the component-wise conditionally unbiased LMMSE (CWCU-LMMSE) estimates of an equivalent linear model with Gaussian priors and noise. Based on the connection between EP and constrained Bethe Free Energy (BFE) optimization, we provide an alternative derivation of the large-system-limit (LSL) BFE for the GLM, which characterizes the stationary points of reGVAMP. Building on the LSL-BFE formulation, we propose the AMBGAMP algorithm, for which convergence can be proven.

In the second part, we investigate message-passing-based algorithms for bilinear estimation. Although orthogonal pilot sequences are assumed, randomly generated pilot sequences can also be accommodated using methods developed in the first part. We first study a hybrid EM-EP framework, in which EP mitigates multi-user interference while EM converts single-user bilinear estimation into linear estimation problems. Owing to the structure of EP, the EM-EP framework can be implemented in a distributed manner by appropriately rewriting the factorization scheme. We also analyze the impact of loops in factor graphs and show through simulations that loop-free factorizations lead to significantly faster convergence.

A simplified EP-based algorithm for bilinear estimation is then developed using LSL approximations. These simplifications enable decentralized implementation.

By analyzing the BFE of the bilinear estimation problem, we propose the VB-EP algorithm, which converts bilinear estimation into linear estimation via VB while preserving dependencies between channel and data variables. This algorithm supports both distributed and decentralized implementations. Compared to EP-based methods, VB-EP reduces computational complexity at the AP side. However, this reduction comes at the cost of additional constraints in the BFE optimization problem, which shrink the feasible set. To address this limitation, we further develop hierarchical EP, which exploits the structure of widely used QAM constellations to reduce AP-side complexity from linear to logarithmic with respect to the constellation size.

We also investigate the GF-CF network, which is the key technology for machine-type communications in IoT and autonomous driving. A distributed iterative maximum likelihood (ML) method is proposed for activity detection, achieving higher accuracy compared to existing message-passing-based approaches.

Finally, we focus on more fundamental problems rooted in EP. Non-integrable beliefs prevents the EP algorithms from proceeding. Most of researchers avoid this problem by limiting the messages to integrable ones. The non-integrable-belief problem becomes more common as the factorization schemes grow more complicated, especially in bilinear joint estimation and detection problems, where we have GMM beliefs. We generalized the problem from the CF communication as estimating the mean and variance of the a random variable whose prior is in the form of GMM product. Two variants of EP are proposed that prevent the non-integrable belief while allowing the non-integrable messages. However, these methods cannot be applied to arbitrary factorization schemes. We further develop a more general framework for EP that prevents the non-integrable beliefs and can be applied in more complicated factorization structures.

13.1 Future Work

By using the concept of equivalent Gaussian priors and noise that preserve posterior statistics, the LSL-BFE in each AMBGAMP iteration can be solved analytically by jointly satisfying the Karush–Kuhn–Tucker (KKT) conditions. Preliminary simulations indicate that this approach outperforms ADMM-based methods. To further reduce computational complexity, particularly the matrix inversion required in least-squares steps, adaptive accelerated optimization methods will be investigated.

The relation between VL-EP and VB-EP is worth further investigating. These two algorithms share a common mechanism which converts the single user bilinear estimation problem to single user linear estimation problems. Since we know the stationary points of VB-EP, it will be interesting if we can find the stationary points of VL-EP.

We find that by combining the ideas of VB-EP and Hierarchical EP, the resulting message passing algorithm contains only integrable beliefs and integrable messages. It will be interesting to compare the simulation results with other algorithms.

The EP framework that prevents non-integrable beliefs can only be applied in sequential order. A new framework that allows parallel update is also worth investigating.

Bibliography

- [1] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, “Cell-Free Massive MIMO Versus Small Cells,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1834–1850, 2017.
- [2] A. Karataev, C. Forsch, and L. Cottatellucci, “Bilinear Expectation Propagation for Distributed Semi-Blind Joint Channel Estimation and Data Detection in Cell-Free Massive MIMO,” *IEEE Open Journal of Signal Processing*, 2024.
- [3] C. Forsch, Z. Zhao, D. Slock, and L. Cottatellucci, “Bayesian Learning for Pilot Decontamination in Cell-Free Massive MIMO,” in *2025 28th International Workshop on Smart Antennas (WSA)*. IEEE, 2025, pp. 19–25.
- [4] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. De Carvalho, “Sparse Signal Processing for Grant-Free Massive Connectivity: A Future Paradigm for Random Access Protocols in the Internet of Things,” *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 88–99, 2018.
- [5] M. J. Wainwright, M. I. Jordan *et al.*, “Graphical Models, Exponential Families, and Variational Inference,” *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, 2008.
- [6] D. Zhang, X. Song, W. Wang, G. Fettweis, and X. Gao, “Unifying Message Passing Algorithms Under the Framework of Constrained Bethe Free Energy Minimization,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4144–4158, 2021.
- [7] T. P. Minka, “Expectation Propagation for Approximate Bayesian Inference,” *arXiv preprint arXiv:1301.2294*, 2013.
- [8] K. Murphy, Y. Weiss, and M. I. Jordan, “Loopy Belief Propagation for Approximate Inference: An Empirical Study,” *arXiv preprint arXiv:1301.6725*, 2013.
- [9] T. P. Minka, “A Family of Algorithms for Approximate Bayesian Inference,” Ph.D. dissertation, Massachusetts Institute of Technology, 2001.
- [10] K.-H. Ngo, M. Guillaud, A. Decurninge, S. Yang, and P. Schniter, “Multi-User Detection Based on Expectation Propagation for the Non-Coherent SIMO Multiple Access Channel,” *IEEE Transactions on Wireless Communications*, 2020.
- [11] T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter, “Approximate Inference Techniques with Expectation Constraints,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 11, 2005.
- [12] Q. Zou and H. Yang, “A Concise Tutorial on Approximate Message Passing,” *arXiv preprint arXiv:2201.07487*, 2022.

- [13] P. Schniter, “A Simple Derivation of AMP and its State Evolution via First-Order Cancellation,” *IEEE Transactions on Signal Processing*, 2020.
- [14] D. L. Donoho, A. Maleki, and A. Montanari, “Message Passing Algorithms for Compressed Sensing,” *Proceedings of the National Academy of Sciences (PNAS)*, vol. 106, no. 45, pp. 18 914–18 919, 2009.
- [15] Q. Zou, H. Zhang, C.-K. Wen, S. Jin, and R. Yu, “Concise Derivation for Generalized Approximate Message Passing Using Expectation Propagation,” *IEEE Signal Processing Letters*, 2018.
- [16] S. Rangan, “Generalized Approximate Message Passing for Estimation with Random Linear Mixing,” in *IEEE International Symposium on Information Theory (ISIT 2011)*. IEEE, 2011, pp. 2168–2172.
- [17] F. Krzakala, A. Manoel, E. W. Tramel, and L. Zdeborová, “Variational Free Energies for Compressed Sensing,” in *IEEE International Symposium on Information Theory (ISIT 2014)*. IEEE, 2014, pp. 1499–1503.
- [18] S. Rangan, P. Schniter, E. Riegler, A. K. Fletcher, and V. Cevher, “Fixed Points of Generalized Approximate Message Passing with Arbitrary Matrices,” *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7464–7474, 2016.
- [19] S. Rangan, P. Schniter, and A. K. Fletcher, “Vector Approximate Message Passing,” *IEEE Transactions on Information Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.
- [20] R. Gholami, L. Cottatellucci, and D. Slock, “Message Passing for a Bayesian Semi-Blind Approach to Cell-Free Massive MIMO,” in *Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2021.
- [21] H. Jiang, X. Yuan, and Q. Guo, “Hybrid Vector Message Passing for Generalized Bilinear Factorization,” *arXiv preprint arXiv:2401.03626*, 2024.
- [22] —, “Generalized Bilinear Factorization via Hybrid Vector Message Passing,” *IEEE Transactions on Signal Processing*, vol. 72, pp. 5675–5690, 2024.
- [23] J. T. Parker, P. Schniter, and V. Cevher, “Bilinear Generalized Approximate Message Passing—Part I: Derivation,” *IEEE Transactions on Signal Processing*, 2014.
- [24] M. Triki and D. T. Slock, “Component-Wise Conditionally Unbiased Bayesian Parameter Estimation: General Concept and Applications to Kalman Filtering and LMMSE Channel Estimation,” in *Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2005, pp. 670–674.
- [25] M. Huemer and O. Lang, “CWCU LMMSE Estimation: Prerequisites and Properties,” *arXiv preprint arXiv:1412.1567*, 2014.
- [26] C. C. Moallemi and B. Van Roy, “Consensus Propagation,” *IEEE Transactions on Information Theory*, 2006.
- [27] K. Takeuchi, “Decentralized Generalized Approximate Message-Passing for Tree-Structured Networks,” *arXiv preprint arXiv:2311.01032*, 2023.
- [28] R. F. Fischer, C. Sippel, and N. Goertz, “VAMP with Vector-Valued Diagonalization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*. IEEE, 2020, pp. 9110–9114.

- [29] J. A. Nelder and R. W. Wedderburn, “Generalized Linear Models,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 135, no. 3, pp. 370–384, 1972.
- [30] J. D. Blume, L. Su, R. M. Olveda, and S. T. McGarvey, “Statistical Evidence for GLM Regression Parameters: A Robust Likelihood Approach,” *Statistics in Medicine*, vol. 26, no. 15, pp. 2919–2936, 2007.
- [31] J. Boutros and G. Caire, “Iterative Multiuser Joint Decoding: Unified Framework and Asymptotic Analysis,” *IEEE Transactions on Information Theory*, vol. 48, no. 7, pp. 1772–1793, 2002.
- [32] Y. Wang, R. Wang, S. S. Du, and A. Krishnamurthy, “Optimism in Reinforcement Learning with Generalized Linear Function Approximation,” *arXiv preprint arXiv:1912.04136*, 2019.
- [33] P. Schniter, S. Rangan, and A. K. Fletcher, “Vector Approximate Message Passing for the Generalized Linear Model,” in *50th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2016, pp. 1525–1529.
- [34] Z. Zhao, F. Xiao, and D. Slock, “Approximate Message Passing for Not So Large Non-IID Generalized Linear Models,” in *2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2023, pp. 386–390.
- [35] T. Minka *et al.*, “Divergence Measures and Message Passing,” Citeseer, Tech. Rep., 2005.
- [36] C. Sippel and R. F. Fischer, “Variants of VAMP for Signal Recovery in Wireless Sensor Networks,” in *2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*. IEEE, 2022, pp. 37–42.
- [37] Z. Zhao, F. Xiao, and D. Slock, “Vector Approximate Message Passing for Not So Large Non-IID Generalized I/O Linear Models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2024)*. IEEE, 2024, pp. 13 281–13 285.
- [38] S. Wagner, R. Couillet, M. Debbah, and D. T. Slock, “Large System Analysis of Linear Precoding in Correlated MISO Broadcast Channels Under Limited Feedback,” *IEEE Transactions on Information Theory*, vol. 58, no. 7, 2012.
- [39] M. E. Tipping, “Sparse Bayesian Learning and the Relevance Vector Machine,” *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.
- [40] D. P. Wipf and B. D. Rao, “Sparse Bayesian Learning for Basis Selection,” *IEEE Transactions on Signal Processing*, vol. 52, no. 8, 2004.
- [41] C. K. Thomas and D. Slock, “Low-Complexity Static and Dynamic Sparse Bayesian Learning Combining BP, VB, and EP Message Passing,” in *Asilomar Conference on Signals, Systems, and Computers*, CA, USA, 2019.
- [42] S. Rangan, A. K. Fletcher, P. Schniter, and U. S. Kamilov, “Inference for Generalized Linear Models via Alternating Directions and Bethe Free Energy Minimization,” *IEEE Transactions on Information Theory*, vol. 63, no. 1, pp. 676–697, 2016.
- [43] Q. Guo and J. Xi, “Approximate Message Passing with Unitary Transformation,” *arXiv preprint arXiv:1504.04799*, 2015.
- [44] D. Slock, “Convergent Approximate Message Passing by Alternating Constrained Minimization of Bethe Free Energy,” in *IEEE Forum on Research and Technologies for Society and Industry (RTSI 2022)*, Paris, France, 2022.

- [45] —, “Convergent Approximate Message Passing,” in *IEEE Mediterranean Conference on Communications and Networking (MeditCom 2022)*, Athens, Greece, 2022.
- [46] C. K. Thomas and D. Slock, “Alternating Constrained Minimization-Based Approximate Message Passing,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, 2023.
- [47] M. Triki and D. Slock, “Investigation of Some Bias and MSE Issues in Block-Component-Wise Conditionally Unbiased LMMSE,” in *Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, USA, 2006.
- [48] S. Wagner, R. Couillet, M. Debbah, and D. T. M. Slock, “Large System Analysis of Linear Precoding in Correlated MISO Broadcast Channels Under Limited Feedback,” *IEEE Transactions on Information Theory*, 2012.
- [49] D. T. Slock, “Nonlinear MMSE Using Linear MMSE Bricks and Application to Compressed Sensing and Adaptive Kalman Filtering,” IEEE ICASSP Expert to Non-Expert (ETON) Primer, 2020.
- [50] R. Gholami, L. Cottatellucci, and D. Slock, “Tackling Pilot Contamination in Cell-Free Massive MIMO by Joint Channel Estimation and Linear Multi-User Detection,” in *IEEE International Symposium on Information Theory (ISIT)*, 2021.
- [51] Z. Zhao and D. Slock, “Semi-Blind Sparse Channel Learning in Cell-Free Massive MIMO: A CRB Analysis,” in *IEEE International Conference on Communications (ICC)*, 2023.
- [52] E. Nayebi and B. D. Rao, “Semi-Blind Channel Estimation for Multiuser Massive MIMO Systems,” *IEEE Transactions on Signal Processing*, 2018.
- [53] K. Mawatwal, D. Sen, and R. Roy, “A Semi-Blind Channel Estimation Algorithm for Massive MIMO Systems,” *IEEE Wireless Communications Letters*, 2017.
- [54] E. Nayebi and B. D. Rao, “Semi-Blind Channel Estimation in Massive MIMO Systems with Different Priors on Data Symbols,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.
- [55] Z. Zhao and D. Slock, “Bilinear Hybrid Expectation Maximization and Expectation Propagation for Semi-Blind Channel Estimation,” in *32nd European Signal Processing Conference (EUSIPCO 2024)*. IEEE, 2024, pp. 2122–2126.
- [56] —, “Decentralized Expectation Propagation for Semi-Blind Channel Estimation in Cell-Free Networks,” *arXiv preprint arXiv:2410*, 2024.
- [57] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, “Cell-Free Massive MIMO: Uniformly Great Service for Everyone,” in *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2015, pp. 201–205.
- [58] Z. Zhao and D. Slock, “Bethe Free Energy and Extrinsic in Approximate Message Passing,” Asilomar Conference on Signals, Systems, and Computers (ASILOMAR), 2023.
- [59] —, “Variance Predictions in VAMP/UAMP with Right Rotationally Invariant Measurement Matrices for Non-IID Generalized Linear Models,” in *31st European Signal Processing Conference (EUSIPCO 2023)*, 2023.
- [60] —, “Improved Variance Predictions in Approximate Message Passing,” in *IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP 2023)*, 2023.

- [61] —, “Decentralized Message Passing for Semi-Blind Channel Estimation in Cell-Free Systems Based on Bethe Free Energy Optimization,” in *2024 58th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2024, pp. 1443–1447.
- [62] C. Forsch, A. Karataev, and L. Cottatellucci, “Distributed Joint User Activity Detection, Channel Estimation, and Data Detection via Expectation Propagation in Cell-Free Massive MIMO,” in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2024.
- [63] J. Goldberger and A. Leshem, “Pseudo-Prior Belief Propagation for Densely Connected Discrete Graphs,” in *IEEE Information Theory Workshop (ITW)*, 2010.
- [64] H. Iimori, T. Takahashi, K. Ishibashi, G. T. F. de Abreu, and W. Yu, “Grant-Free Access via Bilinear Inference for Cell-Free MIMO with Low-Coherence Pilots,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 11, pp. 7694–7710, 2021.
- [65] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier, 2014.
- [66] K. Murphy, Y. Weiss, and M. I. Jordan, “Loopy Belief Propagation for Approximate Inference: An Empirical Study,” 2013. [Online]. Available: <https://arxiv.org/abs/1301.6725>
- [67] A. Ihler, E. Sudderth, W. Freeman, and A. Willsky, “Efficient Multiscale Sampling from Products of Gaussian Mixtures,” *Advances in Neural Information Processing Systems*, vol. 16, 2003.
- [68] E. Sudderth, A. Ihler, W. Freeman, and A. Willsky, “Nonparametric Belief Propagation,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, 2003, pp. I–I.
- [69] K. R. Rayan Ranasinghe, H. L. Senger, G. P. Gonçalves, H. S. Rou, B. S. Chang, G. T. F. de Abreu, and D. Le Ruyet, “Affine Filter Bank Modulation (AFBM): A Novel 6G ISAC Waveform with Low PAPR and OOB,” *arXiv preprint arXiv:2509*, 2025.
- [70] Z. Zhao, F. Xiao, and D. Slock, “Expectations in Expectation Propagation,” *arXiv preprint arXiv:2512.08034*, 2025.
- [71] O. C. Schrempf, O. Feiermann, and U. D. Hanebeck, “Optimal Mixture Approximation of the Product of Mixtures,” in *2005 7th International Conference on Information Fusion*. IEEE, 2005, pp. 8–pp.
- [72] Z. Zhao and D. Slock, “Approximating Univariate Factored Distributions via Message Passing Algorithms,” *arXiv preprint arXiv:2602.01377*, 2026.
- [73] Z. Zhao, J. Chen, and D. Slock, “Avoiding Non-Integrable Beliefs in Expectation Propagation,” *arXiv preprint*, 2026.