Unifying Message Passing Algorithms
Under the Framework of
Constrained Bethe Free Energy Minimization

Dan Zhang, Xiaohang Song, Member, IEEE, Wenjin Wang, Member, IEEE
Gerhard Fettweis, Fellow, IEEE, and Xiqi Gao, Fellow, IEEE

Abstract

Variational message passing (VMP), belief propagation (BP) and expectation propagation (EP) have found their wide applications in complex statistical signal processing problems. In addition to viewing them as a class of algorithms operating on graphical models, this paper unifies them under an optimization framework, namely, Bethe free energy minimization with differently and appropriately imposed constraints. This new perspective in terms of constraint manipulation can offer additional insights on the connection between different message passing algorithms and is valid for a generic statistical model. It also founds a theoretical framework to systematically derive message passing variants. Taking the sparse signal recovery (SSR) problem as an example, a low-complexity EP variant can be obtained by simple constraint reformulation, delivering better estimation performance with lower complexity than the standard EP algorithm. Furthermore, we can resort to the framework for the systematic derivation of hybrid message passing for complex inference tasks. Notably, a hybrid message passing algorithm is exemplarily derived for joint SSR and statistical model learning with near-optimal inference performance and scalable complexity.

Index Terms

Statistical inference, Bethe free energy, message passing algorithms, constrained optimization.

This work has been supported jointly by the German Science Foundation (DFG) and the National Natural Science Foundation of China (NSFC) under the project "Large-Scale and Hierarchical Bayesian Inference for Future Mobile Communication Networks".

D. Zhang was with the Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, Germany, and is currently with Bosch Center for Artificial Intelligence, Germany, Email: dan.zhang2@de.bosch.com. The work was conducted at her stay with the Vodafone Chair.

X. Song and G. Fettweis are with the Vodafone Chair Mobile Communications Systems, Technische Universität Dresden, Germany, Email: {xiaohang.song, gerhard.fettweis}@tu-dresden.de.

W. Wang and X. Gao are with National Mobile Communications Research Laboratory, Southeast University, China, Email: {wangwj, xqgao}@seu.edu.cn.
I. INTRODUCTION

Many signal processing problems can be formulated as a statistical inference task of estimating the latent random variable $x$ given the realization of a statistically related observation $y$. Knowing the likelihood function $p(y|x)$ and the prior density $p(x)$, the inference task, performing under the Bayesian framework, relies on the a-posteriori density $p(x|y) \propto p(y|x)p(x)$. Treating $x$ and $y$ as the input and output of a system, respectively, the a-posteriori density mathematically describes the input-output statistical dependence, permitting inference under various criteria, e.g., minimum mean square error (MMSE) $\hat{x} = E[x|y]$, or maximum a-posteriori (MAP) $\hat{x} = \arg \max_x p(x|y)$.

In practical large-scale wireless communications systems, it might be too complicated to evaluate $p(x|y)$ or to derive its statistical properties such as MMSE and MAP estimates. The reason could be a too large feasible space of $x$, or the form of $p(x|y)$ is analytically intractable. To cope with such cases, we resort to some form of approximations that generally fall into two classes, i.e., deterministic and stochastic approximations. As mentioned in [1], [2], stochastic approximation approaches, e.g., the Markov Chain Monte Carlo, tend to be more computationally demanding. Aiming at low complexity methods for large-scale systems, in this work, we consider a family of deterministic approximations termed variational Bayesian inference.

A. Variational Bayesian inference and message passing

Briefly, variational Bayesian inference attempts to approximate $p(x|y)$ by an alternative density $\hat{b}(x)$. By constraining the form of $\hat{b}(x)$, one can ensure the mathematical tractability of deriving statistical properties on top of it. In the meantime, sufficient approximation accuracy must be ensured to perform inference. To this end, the Kullback-Leibler (KL) divergence (a.k.a. relative entropy), which quantifies the difference between a given density pair, is used here. Limiting to a family $Q$ of densities in the desired form, $\hat{b}(x)$ is chosen to yield the minimal KL divergence with respect to $p(x|y)$ [3]. In the context of physics, such an optimization problem is also known as variational free energy minimization (a.k.a. Gibbs free energy minimization) [4].

There are two well-accepted approaches to construct the specialized density family $Q$. Firstly, the mean field approach defines it as a set of fully factorisable densities, and variational message

---

[1] Here we note that no single approximation technique, neither deterministic nor stochastic, outperforms all others on all problems. In fact, both types of approximations are broad enough research topics to be studied on their own.
passing (VMP) [5] (a.k.a. mean field algorithm) was proposed accordingly as an iterative solution. Here, we note that the expectation maximization (EM) algorithm initially introduced to solve the maximum likelihood (ML) estimation problem [6] can be considered as a special case of VMP. It additionally forces the densities to be Dirac-delta functions with a single parameter.

The second approach constructs $Q$ by exploiting the factorization of $p(x|y)$ or $p(x, y)$ 

2 The variational free energy is then approximated by the so-called Bethe free energy [4], which can be minimized by belief propagation (BP) [7]. However, BP is not well suited to accomplish tasks that involve continuous random variables, e.g., synchronization and channel estimation in communications systems. To tackle this issue, expectation propagation (EP) adds one step to BP, i.e., projecting the beliefs onto a specific function family for analytical tractability [8]. Being the approximate solutions to variational free energy minimization, the above mentioned algorithms have been widely applied to solve problems that arise in communication systems with large dimensions, e.g., multiuser detection [9], [10], coded modulation capacity analysis [11], multiple-input multiple-output (MIMO) channel estimation and data detection [12]–[14].

Approximate message passing (AMP) was developed in [15] for compressed sensing and later generalized by [16] (thereby termed GAMP) for solving the problem in generalized linear systems. They exhibit intrinsic connections to EP in the large system limit, e.g., circumventing matrix inversion in EP with the aid of the self-averaging method [17] or neglecting high-order terms when computing EP messages [18]. Due to the relatively low complexity, they have become pragmatic alternatives to BP and EP for large-scale estimation and detection, e.g., [19]–[23]. Under an i.i.d. MIMO Gaussian channel, the optimality of GAMP for large-scale MIMO detection was assessed in [24].

Large-scale sparse signal recovery (SSR) is relevant to communication systems that are under-determined, e.g., active user detection in massive machine-type communication (mMTC) and channel estimation in millimeter wave broadband communication. Among the existing techniques, one class based on the empirical Bayesian framework is termed sparse Bayesian learning (SBL). Both VMP (including its special case EM) and (G)AMP are applicable [20], [25], [26]. Therefore, they have been applied in the literature for estimating sparse channels, e.g., [27]–[29],

3 Here, the density $p(y)$ involved in $p(x, y) = p(x|y)p(y)$ is treated as a constant as it is not a function of the variable $x$.

3 The EP algorithms in work [17] and [18] are w.r.t. two different types of factorization on the same objective function.
and also for active user detection on random access channels [30].

B. Motivation and contribution of this work

From the above state-of-the-art overview, we notice that most existing theoretical works investigated message passing algorithms individually, even though their heuristic combinations have already found its applications, e.g., [31]–[35]. With a joint use of the mean field and Bethe approaches for approximating the variational free energy, VMP and BP were merged in [36] via partitioning the factors into two corresponding subsets. Apart from that, to our best knowledge, little results have been reported on unifying message passing algorithms in a single mathematical framework. This paper aims at an optimization framework that can link them with a generic statistical model, and the main contributions are summarized as follows:

- We construct the framework based on constrained Bethe free energy minimization. In particular, the Bethe free energy, an approximation to the variational free energy, is used as the objective function. On top of it, we introduce a set of constraint formulation methods such that BP, EP, and VMP can be analytically attributed to corresponding constrained Bethe free energy minimization.
- From this novel perspective of constraint manipulation, we systematically derive new message passing variants, in particular hybrid ones for complex statistical inference problems. It is noted that under our framework, BP and VMP can be combined in a more generalized manner than that in [36]. To further ease the understanding and implementation of hybrid message passing, the conventional factor graph is adapted accordingly for visualization.
- We exemplarily address a classic SSR problem under the developed framework. Through constraint reformulation, we derive new message passing variants with and without perfect knowledge of the statistical model, outperforming multiple benchmark algorithms in both performance and complexity.
- We showcase our framework for solving practical communications problems in systems such as massive MIMO and millimeter-wave communications.

The rest of this paper is organized as follows: Section II introduces the message passing algorithms, i.e., BP, EP and VMP, for variational free energy minimization, respectively. The key part of the paper is given in Section III. It describes the optimization framework that can unify BP, EP and VMP under different constraints. In Section IV a SSR problem is considered
as an example for practicing the developed framework, and four wireless application examples are also provided with this framework being their new viewpoints. Finally, a conclusion and outlook are presented in Section V.

II. VARIATIONAL FREE ENERGY MINIMIZATION

In this section, we link BP, EP and VMP algorithms to variational free energy minimization [4]. The derivation starts with a density $p(x|y)$ constructed from a non-negative function $f(x)$ as

$$
p(x|y) = \frac{1}{Z} f(x) \quad \text{with} \quad Z = \int f(x) dx, \quad (1)
$$

where $dx$ denotes the Lebesgue measure and counting measure respectively for continuous and discrete case. Using a trial density $b(x)$, we define a new function

$$
F(b) \triangleq \int b(x) \ln b(x) dx - \int b(x) \ln f(x) dx = - \ln Z + D[b(x) \parallel p(x|y)], \quad (2)
$$

where $D[\cdot \parallel \cdot]$ stands for the KL divergence between two distributions. In physics, it is known as the variational free energy of the system and $- \ln Z$ is termed Helmholtz free energy. By minimizing $F(b)$ over all possibilities of the trial density $b(x)$, the solution is straightforward, i.e., $\hat{b}(x) = p(x|y)$ and $F(\hat{b}) = - \ln Z$. Therefore, the minimization of $F(b)$ is an exact procedure to compute $- \ln Z$ and recover $p(x|y)$. However, it is not always tractable. One common approximate solution is to limit the feasible set of $b(x)$ to a density family $Q$, i.e.,

$$
\hat{b}(x) = \arg \min_{b(x) \in Q} F(b). \quad (3)
$$

Apparently, the choice of $Q$ determines the fidelity and tractability of the resulting approximation $\hat{b}(x) \approx p(x|y)$ and $F(\hat{b}) \approx - \ln Z$. In the following, we will show how to construct $Q$ such that BP, EP and VMP are iterative solutions to the corresponding minimization problem.

A. Mean field approximation

The mean field approximation to the variational free energy minimization problem [3] constructs $Q$ as a collection of fully factorizable densities, i.e.,

$$
b(x) = \prod_{i} b_i(x_i) \quad \forall \ b(x) \in Q. \quad (4)
$$
With such a family, the primary problem (3) becomes

\[ \{ \hat{b}_i(x_i) \} = \arg \min_{\{ b_i(x_i) \}} \mathcal{F} \left( \prod_i b_i(x_i) \right). \]  

(5)

The problem (5) is convex with respect to each individual \( b_i(x_i) \), but non-convex when jointly considering \( \{ b_i(x_i) \} \). The optimal update equation for \( b_i(x_i) \) with the others being fixed is

\[ b_i(x_i) \propto e^{\int \ln f_a(x_a) \prod_{i' \in I_a \setminus i} b_{i'}(x_{i'}) \mathrm{d} x_{a \setminus i}}, \]  

(6)

where \( I_a \) stands for the index set of the entries of \( x_a \) in the complete vector \( x \) and the integral is with respect to \( x_a \) except \( x_i \). With proper initialization, \( \{ b_i(x_i) \} \) can then be successively and iteratively updated using (6) until a local minimum or a saddle point is reached.

We note that (6) is identical to the message update rule of VMP [5], implying VMP as an iterative solution to (5). Given the full factorization of the trial density \( b(x) \) from the construction of \( Q \) in (4), the statistical dependence between variables is therefore overlooked by VMP for the sake of low-complexity inference.

**B. Bethe approximation**

The Bethe approximation attempts to preserve the local dependence during inference by exploiting the intrinsic factorization of \( f(x) \), namely,

\[ f(x) = \prod_a f_a(x_a), \]  

(7)

where the argument \( x_a \) of the factor function \( f_a(x_a) \) is a subvector of the vector \( x \). We can visualize the factorization by means of factor graph [37]. As the example being depicted in Fig. 1, each entry of \( x \) is depicted as a variable node, while factor nodes denote the factor functions \( \{ f_a \} \). One factor node is connected to a set of variable nodes that are its arguments. The variables that are connected to the same factor node are locally dependent. The Bethe
approximation introduces the auxiliary densities \( \{ b_a(x_a) \} \) to describe their statistical relations. In addition, it defines \( \{ b_i(x_i) \} \) for each variable node. Aiming at the global dependence across multiple factor nodes, the Bethe approximation requires \( \{ b_a(x_a) \} \) and \( \{ b_i(x_i) \} \) to fulfill the marginalization consistency constraint

\[
\int b_a(x_a) \, dx_{a\setminus i} = b_i(x_i) \quad \forall i \forall a \in A_i. \tag{8}
\]

If the factor graph of \( f(x) \) has a tree structure, the optimal setting for \( Q \) is

\[
b(x) = \frac{\prod_a b_a(x_a)}{\prod_i |b_i(x_i)|^{A_i-1}} \quad \forall b(x) \in Q, \tag{9}\]

where \( A_i \) stands for the cardinality of set \( A_i \), and \( A_i \) collects the indices of the factor functions that with \( x_i \) being one of their arguments. Substituting (9) back into (2) and exploiting the fact that \( \{ b_a(x_a), b_i(x_i) \} \) in this case are marginals of \( b(x) \), we obtain the Bethe free energy

\[
F_B(\{b_a\}, \{b_i\}) = \sum_a \int b_a(x_a) \ln \frac{b_a(x_a)}{f_a(x_a)} \, dx_a - \sum_i (A_i - 1) \int b_i(x_i) \ln b_i(x_i) \, dx_i. \tag{10}\]

The optimal solution \( \hat{b}_i(x_i) \) of

\[
\min_{\{b_a\},\{b_i\}} F_B(\{b_a\}, \{b_i\}) \quad \text{s.t. } \tag{11}\]

is exactly equal to \( p(x_i) \) \[^4\] and can be found by BP.

If the factor graph contains cycles, we can still formulate and solve the constrained Bethe free energy minimization problem as given in (11). However, the obtained results are only approximations to \( \{ p(x_i) \} \). The authors of \[^4\] have proven the fixed points of BP satisfy the necessary conditions for being an interior optimum (local minimum or maximum) of (11) in the discrete case. In \[^38\], stable fixed points of BP were shown to be local minima. Therefore, one can regard BP as an iterative solution to the minimization problem in (11).

On top of the factor graph, e.g., Fig. 1, we can describe BP by specifying two rules for computing messages that are: i) from a factor node \( f_a \) to a variable node \( x_i \) and ii) in the reverse direction, which are respectively given as

\[
m_{a \rightarrow i}(x_i) \propto \int f_a(x_a) \prod_{i' \in I_a \setminus i} n_{i' \rightarrow a}(x_{i'}) \, dx_{a \setminus i} \quad \text{and} \quad n_{i \rightarrow a}(x_i) = \prod_{a' \in A_i \setminus a} m_{a' \rightarrow i}(x_i). \tag{12}\]

The integral in computing \( m_{a \rightarrow i}(x_i) \) represents the local marginalization associated to the factor

---

[^4]: Since the normalization and non-negative constraints are default setting for any valid density, we will omit them for brevity.
node \( f_a \) and with respect to \( x_i \). The messages \( \{m_{a\rightarrow i}(x_i), n_{i\rightarrow a}(x_i)\} \) with proper initialization can be iteratively and successively updated, representing a message passing flow on the factor graph \([37]\). After a termination condition is satisfied, the target \( f(x_i) \) is approximated by

\[
b(x_i) \propto m_{a\rightarrow i}(x_i)n_{i\rightarrow a}(x_i) = \prod_{a' \in A_i} m_{a'\rightarrow i}(x_i) \quad \forall a \in A_i.
\]

Here we omit the specification of the normalization terms above as they can be case-dependent and often play a negligible role in computation.

Since the marginalization consistency constraint (8) can often be too complicated to yield tractable messages \( \{m_{a\rightarrow i}(x_i), n_{i\rightarrow a}(x_i)\} \), one natural solution is constraint relaxation, such as simplifying it to moment matching \([39]\), i.e.,

\[
E_{b_i}[t(x_i)] = E_{b_i}[t(x_i)],
\]

where \( t(x_i) \) stands for the sufficient statistics of \( x_i \) that are of concern. As shown in \([39]\), the message update rule of EP can be derived from solving the stationary point equations of the Bethe free energy under the constraint (14) with \( t(x_i) \) from exponential family distributions, e.g., first- and second-order moment matching for Gaussian-like message passing. The message update rules in (12) for BP also apply for EP except the computation of \( m_{a\rightarrow i}(x_i) \) involves one additional projection step

\[
m_{a\rightarrow i}(x_i) \propto \text{Proj}_Q \left( c \int f_a(x_a) \prod_{i' \in I_a} n_{i'\rightarrow a}(x_{i'}) dx_{a \setminus i} \right) / n_{i\rightarrow a}(x_i),
\]

where \( Q \) stands for the exponential family characterized by \( t(x_i) \) and the parameter \( c \) is chosen to make the argument of \( \text{Proj}_Q(\cdot) \) a density of \( x_i \). The advantage of using sufficient statistics from the exponential family is that the computational complexity of integration and multiplication becomes tractable, making EP a pragmatic alternative to BP when the latter becomes intractable.

### III. Bethe Approximation Based Optimization Framework

The viewpoint of both BP and EP aiming to minimize Bethe free energy but under differently formalized constraints inspires us to develop a mathematical framework that permits to systematically derive message passing variants through constraint manipulation. In the following, we first show how to formulate the constraints such that VMP can be analytically attributed to the corresponding constrained Bethe free energy minimization. From a novel perspective of unifying BP, EP and VMP via constraint manipulation, we subsequently derive hybrid message passing
variants in a structured manner. Finally, we introduce a modification onto the conventional factor graph for visualizing hybrid message passing.

A. VMP under constrained Bethe free energy minimization

VMP is an iterative scheme to solve (5). We find that it is actually equivalent to adding the following constraint to the Bethe problem (11) targeted by BP

$$b_a(x_a) = \prod_{i \in \mathcal{I}_a} b_a(x_i) \quad \forall a,$$

(16)

where \( b_a(x_i) \) is a marginal of \( b_a(x_a) \). Formally, the optimization problem (5) with VMP being a usable solution can be alternatively written as Bethe free energy minimization under both the marginalization and factorization constraint

$$\min_{\{b_a\},\{b_i\}} F_B(\{b_a\}, \{b_i\}) \quad \text{s.t. (8), (16)}.$$  

(17)

The additional factorization constraint in (16) trivializes the marginalization consistency constraints of BP, i.e., (8), and reduces the Bethe free energy to the objective function in (5). Furthermore, such factorization removes the local dependence captured by \( b_a(x_a) \), thereby turning BP into VMP.

B. Hybrid VMP-BP

Instead of fully factorizing \( b_a(x_a) \), we can selectively ignoring the correlation of subsets through partial factorization

$$b_a(x_a) = \prod_v b_{a,v}(x_{a,v}),$$

(18)

where \( x_{a,v} \) is a subvector of \( x_a \), and its entries in the complete vector \( x \) are recorded by the index set \( \mathcal{I}_{a,v} \subseteq \mathcal{I}_a \) that is mutually disjoint w.r.t. index sets of other subvectors of \( x_a \), i.e., \( \mathcal{I}_{a,v} \cap \mathcal{I}_{a,v'} = \emptyset \). Through such constraint manipulation, we reach to a new optimization problem

$$\min_{\{b_a\},\{b_i\}} F_B(\{b_a\}, \{b_i\}) \quad \text{s.t. (8) and (18)}.$$  

(19)

As (18) retains partial local dependence among the elements of \( x_{a,v} \) and we are interested in the beliefs of each element, we expect that the iterative solution to (19) will exploit the synergy of BP and VMP. We refer to the Appendix A for the detailed derivation and the resulting message
passing rules are summarized below

\[ m_{a\rightarrow i}(x_i) \propto \int \text{d}x_{a,v} \prod_{i' \in I_a \cap i} n_{i'\rightarrow a}(x_{i'}) \cdot \left[ e \int \Pi_{\nu' \neq \nu} b_{a,\nu'}(x_{a,\nu'}) \ln f_a(x_a) \text{d}x_{a,v} \right] \quad (20) \]

\[ n_{i \rightarrow a}(x_i) = \prod_{a' \in A \backslash a} m_{a'\rightarrow i}(x_i). \quad (21) \]

The equation (20) and (21) have an identical form to the message update rules of BP given in (12) if we treat

\[ e \int \Pi_{\nu' \neq \nu} b_{a,\nu'}(x_{a,\nu'}) \ln f_a(x_a) \text{d}x_{a,v} \quad (22) \]

as the factor function \( f_a(x_{a,v}) \) of \( x_{a,v} \). Interestingly, the computation of the term in (22) essentially follows the rule of VMP given in (6). So, VMP aims at disentangling \( \{x_{a,v}\} \) and BP is responsible for computing the marginal beliefs with respect to each element in \( \{x_{a,v}\} \). This observation suggests the derived iterative solution being a hybrid VMP-BP.

Taking the factor node \( f_1 \) in Fig. 1 as an example, we factorize its associated belief \( b_1(x_1, x_2, x_3) \) as \( b_1(x_1)b_1(x_2, x_3) \). Following (20) and (21), the inner integral disentangles \( x_1 \) and \( (x_2, x_3) \) and then \( m_{a=1\rightarrow i=1}(x_1) \) boils down to VMP (the outer integral becomes trivial), whereas \( x_2 \) and \( x_3 \) are updated through BP (the outer integral removes the influence from one to the other through marginalization). From this simple example, we can see the hybrid VMP-BP permits the messages departing from the same factor node to follow different message update rules. This is a key difference to [36] that attempts to combine VMP and BP through factor graph partitioning. From their problem formulation and also the application, the messages from the same factor node obey the same message updating rule. For factor nodes belonging to the BP class \( A_{BP} \), the marginalization constraints are applied to variable nodes connected to them, i.e., [36, Equation (19)]. While for factor nodes of the mean field class \( A_{MF} \), their beliefs are fully factorized to that of their variable nodes, making the marginalization consistency constraints trivial to fulfill. One outcome of such formulation is all messages departing from the factor nodes in \( A_{BP}(A_{MF}) \) must follow BP(VMP) rule.

C. Hybrid VMP-BP-EP

Very often when the variables are continuous, the marginalization consistency constraints become intractable and this issue cannot be addressed by adding partial factorization constraints.
The pragmatic idea behind EP is to relax marginalization consistency into weaker moment matching constraints. In this part, we therefore apply this constraint relaxation idea onto the problem (19) for exploiting the synergy of VMP, BP and EP. In other words, we resort to constraint manipulation, i.e., partial factorization and constraint relaxation, for easing the classic constrained Bethe free energy minimization problem.

Specifically, we divide the variable index set $\mathcal{I}$ into two disjoint index subsets, i.e., $\mathcal{I}^E \cap \mathcal{I}^B = \emptyset$ and $\mathcal{I}^E \cup \mathcal{I}^B = \mathcal{I}$. The variables under one subset $\mathcal{I}^B$ are under the marginalization consistency constraints

$$b_{a,v}(x_i) = b_i(x_i) \quad \forall i \in \mathcal{I}^B \forall a \in \mathcal{A}_i. \quad (23)$$

While, for others belong to $\mathcal{I}^E$, their marginalization consistency constraints are relaxed into

$$E_{b_{a,v}}[t_{a,i}(x_i)] = E_{b_i}[t_{a,i}(x_i)] \quad \forall i \in \mathcal{I}^E \forall a \in \mathcal{A}_i, \quad (24)$$

where $v(i)$ indexes the factor of $b_{a,v}(x_a)$ in $(18)$ that contains $x_i$. It is noted that the sufficient statistics $t_{a,i}$ can vary over the factor-variable node pair, i.e., being edge-dependent. This allows different EP variants propagating along different edges. In next Sec. IV we will show even under the same sufficient statistics it would be interesting to exploit different formulations for an improved message passing performance, i.e., first- and second-order moment matching vs. mean-variance matching.

Now, our target problem becomes

$$\min_{\{b_{a,v}\},\{b_i\}} F_B(\{b_{a,v}\},\{b_i\}) \quad \text{s.t.} \quad (24) \quad \text{and} \quad (23) \quad (25)$$

and see Appendix B for deriving a hybrid VMP-BP-EP to solve it. The obtained message update rules are expressed as

$$m_{a\rightarrow i}(x_i) \propto \int \prod_{i' \in \mathcal{I}_a,v \setminus i} n_{i'\rightarrow a}(x_{i'}) \cdot e^{\int \prod_{v' \neq v} b_{a,v'}(x_{a,v'}) \ln f_a(x_a) dx_{a,v}} \quad i \in \mathcal{I}^B \quad (26)$$

$$m_{a\rightarrow i}(x_i) \propto \frac{1}{n_{i\rightarrow a}(x_i)} \text{Proj}_{Q_{a,i}} \left[c \int \prod_{i' \in \mathcal{I}_a,v} n_{i'\rightarrow a}(x_{i'}) \cdot e^{\int \prod_{v' \neq v} b_{a,v'}(x_{a,v'}) \ln f_a(x_a) dx_{a,v}} \right] \quad i \in \mathcal{I}^E \quad (27)$$

$$n_{i\rightarrow a}(x_i) = \prod_{a' \rightarrow \mathcal{A}_i \setminus a} m_{a'\rightarrow i}(x_i) \quad (28)$$

with $c$ being a normalization constant. Comparing with (20) and (21) (i.e., hybrid VMP-BP),
we only have one additional case for the message \( m_{a \rightarrow i}(x_i) \) if \( i \in \mathcal{I}[E] \), i.e., (27). This is mainly because for some variables the marginalization consistency constraints are relaxed to sufficient statistics matching, following the message update rule of EP.

**D. Visualization of hybrid message passing**

To ease the understanding and implementation of the above-derived hybrid message passing, we propose a modification onto the factor graph, see an illustration in Fig. 2. Its key difference to the conventional factor graph, e.g., Fig. 1 is a new type of node termed hyper-variable node. They are associated to the subvectors \( \{x_{a,v}\} \) that appear in the factorization constraints for \( \{b_a(x_a)\} \), e.g., Fig. 2. Each of them is connected to one and only one factor node.

The message update rule for outgoing from a factor node to a hyper-variable node follows the rule of VMP, e.g.,

\[
m'_{a \rightarrow (a,v)}(x_{a,v}) \propto e^\int \ln f_a(x_a) \prod_{v' \neq v} q_{a,v'}(x_{a,v'}) dx_{a,v'}
\]

with \( \{q_{a,v'}\} \) given as

\[
q_{a,v'}(x_{a,v'}) = m'_{a \rightarrow (a,v')}(x_{a,v'}) \prod_{i \in \mathcal{I}_{a,v'}} n_{i \rightarrow (a,v')}(x_i).
\]

We can interpret \( q_{a,v}(x_{a,v}) \) as the belief associated to the hyper-variable node \( x_{a,v} \). It combines the inputs from all neighboring nodes. In the language of VMP, it is also the message going to the factor node from the hyper-variable node.

From a hyper-variable node to a normal variable node and its reverse direction, there are two cases of message updating. If the destination variable node is under the marginalization consistency constraint (i.e., \( x_i \) with \( i \in \mathcal{I}[B] \)), we take the message update rule of BP by treating...
the hyper-variable node as a factor node and taking the message $m'_{a \rightarrow (a,v)}(x_{a,v})$ from the uniquely connected factor node as the factor function

$$m'_{(a,v) \rightarrow i}(x_i) \propto \int m'_{a \rightarrow (a,v)}(x_{a,v}) \cdot \prod_{i' \in \mathcal{I}_{a,v} \setminus i} n'_{i' \rightarrow (a,v)}(x_{i'}) dx_{a,v \setminus i},$$  \hspace{1cm} (31)

where $n'_{i \rightarrow (a,v)}(x_i)$ given as

$$n'_{i \rightarrow (a,v)}(x_i) = \prod_{a' \in \mathcal{A} \setminus a} m'_{(a',v(i)) \rightarrow i}(x_i)$$  \hspace{1cm} (32)

corresponds to the message from the variable node to the hyper-variable node following the BP rule. As compared to the previously derived hybrid VMP-BP, substituting (29) into (31) let $m'_{(a,v) \rightarrow i}(x_i)$ become equivalent to $m_{a \rightarrow i}(x_i)$ given in (20). Additionally, we have the associations $n'_{i \rightarrow (a,v)}(x_i) \leftrightarrow n_{i \rightarrow a}(x_i)$ and $b_{a,v}(x_{a,v}) \leftrightarrow q_{a,v}(x_{a,v})$.

In the other case where the destination variable node is under the moment matching constraint (i.e., $x_i$ with $i \in \mathcal{I}_{[E]}$), we shall switch to the EP rule, namely

$$m'_{(a,v) \rightarrow i}(x_i) \propto \frac{1}{n'_{i \rightarrow (a,v)}(x_i)} \text{Proj}_q \left[ c \int m'_{a \rightarrow (a,v)}(x_{a,v}) \cdot \prod_{i' \in \mathcal{I}_{a,v}} n'_{i' \rightarrow (a,v)}(x_{i'}) dx_{a,v \setminus i} \right],$$  \hspace{1cm} (33)

where $n'_{i' \rightarrow (a,v)}(x_{i'})$ as the message in the reverse direction still follows (32). Comparing with the former case, the additional $m$-projection step is the only difference here. This coincides with the known difference between EP and BP.

It is worth noting that, comparing (31), (33), (32) with (26), (27), (28), those equations have only slightly differences on notations. This is because $\{\mathcal{I}_{a,v}\}$ are mutually disjoint, and $m_{a \rightarrow i}(x_i)$, $n_{i \rightarrow a}(x_i)$ in Sec. III-B and Sec. III-C are actually the messages passed by the hyper-variable nodes $\{x_{a,v}\}$, i.e., denoting as $m'_{(a,v) \rightarrow i}(x_i)$ and $n'_{i' \rightarrow (a,v)}(x_{i'})$ in this part, respectively.

In short, our framework specifies the message update rules between different types of nodes on this modified factor graph, namely, defining the algorithmic structure. On top of this structure, scheduling remains as a design freedom. In principle, the order of message updating and propagating on the modified factor graph can be arbitrary, depending on applications, and may lead to different results and convergence behaviors.

E. Summary

In this section, based the framework of constrained Bethe free energy minimization, we unify three widely used message passing algorithms for a generic model. Under the same objective
function (i.e., the Bethe free energy), BP, EP and VMP were associated to the marginalization consistency, moment matching and partial factorization constraints, respectively. Moment matching is weaker than marginalization consistency, but beneficial to limit the form of messages. Partial factorization permitted to ignore the variable dependencies to certain extent, easing local marginalization at the factor nodes \( \{f_a(x_a)\} \). Therefore, we can interpret moment matching and partial factorization as constraint manipulation methods to trade inference fidelity for tractability.

Combining three types of constraints in a general manner, systematic derivations led us to hybrid VMP-BP-EP variants. We further mapped the corresponding message passing procedures onto a modified factor graph, visualizing hybrid message passing to assist practical uses.

IV. Application Example

This section aims at the application of the developed framework. We start from a SSR problem, which relates to an active user data estimation problem in massive machine-type communication (mMTC). First, assuming perfect knowledge of the statistical model, we systematically derive a new message passing algorithm that is an outcome of manipulating the constraints for deriving EP. Interestingly, it exhibits high similarity with (G)AMP, thereby being implementation friendly for large-scale systems. Next, we present a systematic derivation of hybrid message passing to efficiently solve the SSR problem without perfectly knowing the statistical model. At the end of this section, we showcase how to apply our framework for solving other practical communications problems such as massive MIMO detection and millimeter-waver channel estimation.

A. EP and its variant for sparse signal recovery (SSR)

The mMTC system is modeled with a single access point of \( N \) antennas and \( M \) users of single antenna each. We write the received signal in up-link as

\[
y = Hx + w, \tag{34}
\]

where the channel matrix \( H \in \mathbb{C}^{N \times M} \) consisting of \( N \) rows \( \{h_n\} \) represents a linear transform on the vector representation \( x \) of the transmitted signals from all users, and the noise vector \( w \) has i.i.d. entries following the Gaussian distribution \( \mathcal{CN}(w_n; 0, \lambda^{-1}) \) where \( w_n \) denotes the \( n \)-th entry of \( w \). With i.i.d. Gaussian channel model, the entries of \( H \) follows distribution \( \mathcal{CN}(h_{nm}; 0, N^{-1}) \). The general goal here is to estimate \( x \in \mathbb{C}^M \) based on the observation
vector $y \in \mathbb{C}^N$ and the knowledge of $H$. Typically, only a few users among hundreds of them is active, making estimating $x$ as a SSR problem. In the context of SSR, the unknown vector $x$ represents a sparse signal only with a few non-zero entries. To simplify the discussion, we assume Gaussian signaling for those non-zero entries. Further assuming independent data symbols at different users, the vector $x$ follows a Bernoulli-Gaussian distribution, i.e., $p(x) = \prod_m (1 - \rho) \delta(x_m) + \rho \mathcal{CN}(x_m; 0, \alpha_m)$, where $\alpha_m$ models the signal power of user $m$ if it is active.

This prior knowledge is critical to reliably estimate $x$ in particular when the system is large-scale and underdetermined ($N \ll M$).

We note that though the considered system model in (34) is linear and additive Gaussian, the following derivation is straightforwardly extendable for generalized linear systems, i.e., the conditional density $p(y | Hx)$ being arbitrary.

1) Formulation of Bethe free energy: When the a-priori density $p(x)$ of $x$ and the noise variance $\lambda^{-1}$ are known, our goal is to estimate $\{x_m\}$ by computing the marginals $\{p(x_m | y; H, \lambda)\}$ of $p(x | y; H, \lambda)$. BP is not a good choice in this continuous case. Alternatively, we consider EP under the first- and second-order moment matching constraints, yielding the messages in the Gaussian family with the sufficient statistic $t(x_i) = [\text{Re}(x_i), \text{Im}(x_i), |x_i|^2]^T$.

If taking the straightforward factorization $p(x | y; H, \lambda) \propto p(x)p(y | x; H, \lambda)$, the resulting EP algorithm will require matrix inversion with complexity $O(N^2 M)$, e.g., like S-AMP in [17].

Considering the complexity issue of a large-scale system, we introduce an auxiliary vector $z \in \mathbb{C}^N$ with the relation $z = Hx$. The target marginal $p(x_m | y; H, \lambda)$ is then alternatively proportional to the outcome of marginalizing

$$f(z, x) = p(y | z; \lambda)p(x) \prod_{n=1}^{N} \delta(z_n - h_n x)$$

with respect to $x_m$. In the following, EP and its variant will be derived with respect to $f(z, x)$.

To this end, we introduce $b_x(z), b_x(x), \{b_{x,z,n}(x, z_n)\}$ and $\{b_{x,m}(x_m), b_{z,n}(z_n)\}$ in accordance with the above factorization. It is noted that the factor function $\delta(z_n - h_n x)$ associated to $b_{x,z,n}(x, z_n)$ is a Dirac delta function. Their KL divergence becomes infinity if there exists a

---

5Since the factor graph of $p(x)p(y | x; H, \lambda)$ is a tree, the corresponding EP is convergent and exactly yields the MMSE estimate of $x$.

6By taking $z$ as one argument of $f(\cdot)$, the following constrained Bethe free energy minimizations will yield estimates of $\{z_n\}$ as well, even though they are not our primary goal.
support set of \( b_{x,z,n}(x, z_n) \) is not a support set of \( \delta(z_n - h_n x) \), i.e.,

\[
\int \int b_{x,z,n}(x, z_n) \ln \frac{b_{x,z,n}(x, z_n)}{\delta(z_n - h_n x)} \, dx \, dz_n = \begin{cases} 
\infty & \text{if } b_{x,z,n}(z_n | x) \neq \delta(z_n - h_n x) \\
-H(b_{x,n}) & \text{else}
\end{cases}
\]  

(36)

where \( b_{x,z,n}(z_n | x) \) and \( b_{x,n}(x) \) are the conditional and marginal density respectively derived from \( b_{x,z,n}(x, z_n) \) according to the Bayes rule, and \( H(\cdot) \) stands for the entropy function. As the minimum of \( F_B(b) \) is of interest, it is evident to let \( b_{x,z,n}(z_n, x) = b_{x,n}(x) \delta(z_n - h_n x) \) so that the Bethe free energy \( F_B(b) \) can be expressed as

\[
F_B(b) = D \left[ b_x(x) \| p(x) \right] + \int b_x(z) \ln \frac{b_z(z)}{p(y | z; \lambda)} \, dz + \sum_{m=1}^M \sum_{n=1}^N [H(b_{z,n}) - H(b_{x,n})].
\]  

(37)

Conventionally, \( F_B(b) \) is minimized under the marginalization consistency constraints given as

\[
\forall m \forall n \quad b_{x,m}(x_m) = \begin{cases} 
\int_{x \setminus m} b_{x,n}(x) \, dx \setminus m & \text{ (a)} \\
\int_{x \setminus m} b_x(x) \, dx \setminus m & \text{ (b)}
\end{cases}
\]  

(38)

\[
\forall n \quad b_{z,n}(z_n) = \begin{cases} 
\int_{x} b_{x,n}(x) \delta(z_n - h_n x) \, dx \\
\int_{z \setminus n} b_z(z) \, dz \setminus n
\end{cases}
\]  

(39)

2) **First- and second-order moment matching:** With a continuous random vector \( x \) of large dimension, the above constraints are often intractable. For the sake of complexity, here we relax them into the first- and second-order moment matching ones

\[
\forall m \forall n \quad E[x_m | b_{x,m}] = E[x_m | b_{x,n}] = E[x_m | b_x],
\]  

(40)

\[
\forall n \quad E[z_n | b_{z,n}] = E[h_n x | b_{z,n}] = E[z_n | b_z],
\]  

\[
\forall m \forall n \quad E[|x_m|^2 | b_{x,m}] = E[|x_m|^2 | b_{x,n}] = E[|x_m|^2 | b_x],
\]  

(41)

Following the method of Lagrange multipliers to minimize \( F_B(b) \) under the constraints (40) and (41), this yields Alg. 1.

Two remarks on Alg. 1 are made as follows: First, it follows the message update rules of EP, but appears differently in comparison with the general EP presentation in terms of \( \{ m_{a \rightarrow i}(x_i), n_{i \rightarrow a}(x_i) \} \). This is an outcome of applying proper simplification with respect to the
specific system model. For instance, as our ultimate goal is to estimate \( x \) the message updates for the auxiliary variable \( z \) are absorbed into the messages for \( x \) in the presentation of Alg. 1. Second, we note that from step 4 to 6, the computations for \( n = 1, 2, \ldots, N \) are simultaneously executed. Such parallel scheduling is beneficial to limit the processing latency when \( N \) tends to be large. In principle, other scheduling schemes are applicable as well. The developed framework permits to define the structure of the algorithm. Scheduling on top of it remains as a design freedom, which is beyond the scope of this work.

3) Mean and variance consistency: In the sequel, we illustrate how to derive a low complexity EP variant by simple constraint reformulation. Namely, under the first-order moment (mean) matching, we can equivalently and alternatively turn the second-order moment matching into the

**Algorithm 1** to minimize \( F_B(b) \) under (40) and (41)

1: Initialization:
   \( b_\phi(x) = p(x) \);
   \( \forall m \in \{1, 2, \ldots, M\}, \alpha_{0,m} = \tau_{0,m} = 0 \);
   \( \forall n \in \{0, 1, \ldots, N\}, \forall m \in \{1, 2, \ldots, M\}, \tilde{\alpha}_{n,m} = \tilde{\tau}_{n,m} = 0 \).
2: repeat
3:   \( \forall m, \)
   \( \tilde{\alpha}_{0,m} = \frac{E[x_m b_\phi]}{\text{Var}[x_m b_\phi]} - \alpha_{0,m} \)
   \( \tilde{\tau}_{0,m} = \frac{1}{\text{Var}[x_m b_\phi]} - \tau_{0,m} \)
4:   For \( n = 1, 2, \ldots, N, \forall m, \)
   \( \alpha_{n,m} = \sum_{n' = 0, n' \neq n}^N \tilde{\alpha}_{n',m} \)
   \( \tau_{n,m} = \sum_{n' = 0, n' \neq n}^N \tilde{\tau}_{n',m} \)
5:   For \( n = 1, 2, \ldots, N, \forall m, \)
   \( \tilde{\alpha}_{n,m} = \frac{h_{nm}^* (y_n - \mu_{zn}) + \alpha_{n,m} |h_{nm}|^2 / \tau_{n,m}}{|h_{nm}|^2} \)
   \( \tilde{\tau}_{n,m} = \frac{\lambda^{-1} + \gamma_n - |h_{nm}|^2 / \tau_{n,m}}{|h_{nm}|^2} \)
6:   For \( n = 1, 2, \ldots, N, \forall m, \)
7:   \( \forall m, \)
   \( \alpha_{0,m} = \sum_{n' = 1}^N \tilde{\alpha}_{n',m} \)
   \( \tau_{0,m} = \sum_{n' = 1}^N \tilde{\tau}_{n',m} \)
8:   \( b_\phi(x) \propto p(x) \prod_m e^{-2Re[\alpha_{n,m} x_m] + \tau_{0,m} |x_m|^2} \)
9: until the termination condition is fulfilled.
10: \( \hat{x}_m = E[x_m | b_\phi, m] \)

**Algorithm 2** to minimize \( F_B(b) \) under (40) and (42)

1: Initialization:
   \( b_\phi(x) = p(x) \);
   \( \forall m \in \{1, 2, \ldots, M\}, \tau_{0,m} = 0 \);
   \( \forall n \in \{0, 1, \ldots, N\}, \beta_n = 0 \);
   \( \forall n \in \{0, 1, \ldots, N\}, \forall m \in \{1, 2, \ldots, M\}, \tilde{\tau}_{n,m} = 0 \).
2: repeat
3:   \( \forall m, \tilde{\tau}_{0,m} = \frac{1}{\text{Var}[x_m b_\phi]} - \tau_{0,m} \)
4:   For \( n = 1, 2, \ldots, N, \forall m, \)
   \( \tau_{n,m} = \sum_{n' = 0, n' \neq n}^N \tilde{\tau}_{n',m} \)
5:   For \( n = 1, 2, \ldots, N, \forall m, \)
   \( \gamma_z,n = \sum_m |h_{nm}|^2 \)
6:   For \( n = 1, 2, \ldots, N, \forall m, \)
    \( \mu_{zn} = h_{nm} E[x | b_\phi] - \beta_n \gamma_z,n \)
7:   For \( n = 1, 2, \ldots, N, \forall m, \)
   \( \beta_n = (y_n - \mu_{zn}) / (\lambda^{-1} + \gamma_z,n) \)
8:   For \( n = 1, 2, \ldots, N, \forall m, \)
   \( \tilde{\tau}_{n,m} = \frac{\gamma_z,n + \lambda^{-1} |h_{nm}|^2}{\tau_{n,m}} \)
9:   \( \forall m, \)
   \( \tau_{0,m} = \sum_{n' = 1}^N \tilde{\tau}_{n',m} \)
10: \( \forall m, \)
   \( \mu_{zn} = E[x_m b_\phi] + \tau_{0,m}^{-1} \sum_{n = 1}^N h_{nm}^* \beta_n \)
11: \( b_\phi(x) \propto p(x) \prod_m C N(x_m; \mu_{zn}, \tau_{0,m}^{-1}) \)
12: until the termination condition is fulfilled.
13: \( \hat{x}_m \leftarrow E[x_m | b_\phi] \)
Algorithm 3 AMP for sparse signal recovery

1: Initialization: \( b_\kappa(x) = p(x), \forall n, \beta_n = 0 \)
2: repeat
3: \( \forall m, \bar{\tau}_{0,m} = \frac{1}{\text{Var}(x_m|b_\kappa)} \)
4: For \( n = 1, 2, \ldots, N \),
5: \( \gamma_{x,n} = \sum_{m=1}^{M} |h_{n,m}|^2 / \bar{\tau}_{0,m} \)
6: For \( n = 1, 2, \ldots, N \), \( \mu_{x,n} = h_n E(x|b_\kappa) - \beta_n \gamma_{x,n} \)
7: \( \beta_n = (y_n - \mu_{x,n}) / (\lambda^{-1} + \gamma_{x,n}) \)
8: \( \forall m, \tau_{0,m} = \sum_{n=1}^{N} \bar{\tau}_{n,m} \)
9: \( \forall m, \mu_{x,m} = E[x_m|b_\kappa] + \tau_{0,m}^{-1} \sum_{n=1}^{N} h_{n,m} \beta_n \)
10: \( b_\kappa(x) \propto p(x) \prod_{m=1}^{M} \mathcal{CN}(x_m; \mu_{x,m}, \tau_{0,m}) \)
11: until the termination condition is fulfilled.
12: \( \hat{x}_m = E[x_m|b_\kappa] \)

Algorithm 4 Hybrid EM-VMP-Alg. 2 for SBL

1: Initialization:
\( \forall m \in \{1, 2, \ldots, M\}, \tau_{0,m} = 0, \alpha_m = 1/M; \)
\( \forall n \in \{0, 1, \ldots, N\}, \beta_n = 0; \)
\( \forall n \in \{0, 1, \ldots, N\}, \forall m\in\{1, 2, \ldots, M\}, \tilde{\tau}_{n,m} = 0; \)
\( b_\kappa(x) = \prod_{m} \mathcal{CN}(x_m; 0, \alpha_m); \)
\( \lambda = \frac{100}{\text{Var}(y)}, c = d = 0, \epsilon = 1.5, \eta = 1, e^{\text{old}} = 0. \)
2: repeat
3: \( p(x) \leftarrow \prod_{m=1}^{M} \mathcal{CN}(x_m; 0, \tilde{\alpha}_m) \)
4: step 3~11 of Alg. 2
5: \( e^{\text{new}} \leftarrow \frac{1}{N}\|y - H\hat{x}\|^2 \)
6: \( \lambda \leftarrow \left(e^{\text{new}} + \frac{1}{N} \sum_{n=1}^{N} \bar{\tau}_{n,m} \right)^{-1} \)
7: \( \tilde{\alpha}_m \leftarrow \{e^{-2+\sqrt{(e^{-2+2+4\eta E(x_m^2|b_\kappa})/(2\eta)}} \}
8: If |e^{\text{new}} - e^{\text{old}}| < 10^{-6} \) and the support of \( E(x|b_\kappa) \) is not reducing, \( \epsilon \leftarrow 0.95\epsilon \)
9: \( e^{\text{old}} \leftarrow e^{\text{new}} \)
10: until the termination condition is fulfilled.

variance consistency constraint, i.e., replacing (41) by

\[
\forall m \forall n \text{ Var}(x_m|b_\kappa) = \text{Var}(x_m|b_\kappa,n) = \text{Var}(x_m|b_\kappa) \quad \text{(42)}
\]

By analogy, we apply the method of Lagrange multipliers for minimizing \( F_B(b) \) under (40) and (42). This leads to Alg. 2. Comparing it with Alg. 1, the updates for \( \{\tilde{\alpha}_{m,n}, \alpha_{m,n}\} \) are avoided, thereby requiring less computation efforts. Both algorithms aim at the same optimization problem as the constraints (40) and (42) are equivalent to those in (40) and (41).

4) Performance comparison: In this part, we compare three message passing algorithms, i.e., Alg. 1, Alg. 2, and AMP [15] (Alg. 3), for SSR in a 250 × 500 linear system, see Fig. 3.

Here, we select AMP as a benchmark algorithm due to its excellent performance attainable with low complexity in the considered system. It is shown to approximate EP (i.e., Alg. 1) in the large system limit, where the high-order terms in the message update equations are ignored [18]. It is also equivalent to S-AMP presented in [17] when \( H \) follows an i.i.d. zero mean Gaussian distribution. We note that S-AMP was derived from free energy minimization under the constraint of first- and second-order moment matching, but it considered the factorization \( p(x)p(y|x; H, \lambda) \)
without using the auxiliary variable $z$. To deal with the resulting matrix inversion, it explicitly exploited the large-scale property of $H$.

The derivations of both Alg. 1 and Alg. 2 on the other hand, are independent of the system dimension. The key difference between Alg. 2 and Alg. 3 (AMP) lies in the computation of $\{\tilde{\tau}_{n,m}\}$. Apart from that, they are nearly identical to each other, being composed of basic arithmetic operations. GAMP [16] generalizes AMP by adapting the step 6 according to an arbitrarily given likelihood $p(y|z)$. By analogy, we simply need to adapt the step 7 in Alg. 2 such that the resulting Alg. 2 is usable for generalized linear models as well.

Fig. 3 (a) shows both AMP and our proposed message passing variant, i.e., Alg. 2 achieve the best performance. On the contrary, the performance of EP Alg. 1 degrades severely when the sparsity ratio $\rho$ is beyond 0.25. From our analysis, such performance degradation mainly arises from the iteration divergence under parallel message updating. This is an interesting observation, indicating that the form of constraints not only impacts the complexity and performance, but also the convergence behavior of message passing, see Fig. 3 (b). Therefore, it is worth to treat constraint manipulation as an important design freedom in the optimization framework when developing practical and effective message passing algorithms for various applications.

To alleviate this issue of Alg. 1 here we empirically introduce damping onto the step 6 of
Alg. 1, i.e.,
\[
\tilde{\alpha}_{n,m} = (1 - \kappa)\tilde{\alpha}_{n,m} + \kappa \frac{h_{nm}^* (y_n - \mu_{s,n}) + \alpha_{n,m} |h_{nm}|^2 / \tau_{n,m}}{\lambda^{-1+\gamma_{z,n}} - |h_{nm}|^2 / \tau_{n,m}} \\
\tilde{\tau}_{n,m} = (1 - \kappa)\tilde{\tau}_{n,m} + \kappa \frac{|h_{nm}|^2}{\lambda^{-1+\gamma_{z,n}} - |h_{nm}|^2 / \tau_{n,m}}
\]
where \( \kappa \in (0, 1] \) is the damping factor. In doing so the damped Alg. 1 is able to deliver similar performance as the others. Meanwhile, it is worth noting that the Alg. 1 to 3 converge at similar speeds, and their complexities per iteration are all in the same order of \( \mathcal{O}(MN) \).

B. Hybrid message passing for joint parameter and statistical model estimation

For practical systems, the prior knowledge of \( \lambda \) and the a-priori density \( p(x) \) are in general not available, making the above-mentioned algorithms not straightforwardly applicable to estimate \( \{x_m\} \). Considering the application example, the signal power of different users can be very different due to different transmit power and large-scale fading conditions, and the user activity probability \( \rho \) is also in general unknown. To overcome the lack of priors, one can build models for \( p(\lambda) \) and \( p(x) \) such that the priors are approximately known up to certain unknown hyperparameters. By jointly estimate \( x \) and those hyperparameters, the SSR problem is extended to an empirical Bayesian estimation problem.

Before including hierarchical prior models of \( \lambda \) and \( x \) into the construction of the Bethe problem, we first model the hyperprior probability density functions of signal power and the noise power as two different Gamma distributions
\[
p(\alpha_m) = \text{Ga}(\alpha_m | \epsilon, \eta); \quad p(\lambda) = \text{Ga}(\lambda | c = 0, d = 0),
\]
where \( (\epsilon, \eta) \) and \( (c, d) \) are the pre-selected shape and rate parameters of the two Gamma distributions, respectively. Next, based on the results in the work [40] that compares different prior models for the estimation of complex sparse signals, we choose the following parameterized Gaussian prior on \( x \)
\[
p(x; \hat{\alpha}) = \prod_{m=1}^{M} \text{CN}(x_m; 0, \hat{\alpha}_m),
\]
where \( \hat{\alpha} = \{\hat{\alpha}_m\} \) denotes the estimates on hyperparameters \( \{\alpha_m\} \).

\(^7\)For the curious readers, a more detailed analysis would show that Alg. 2 and Alg. 3 require approximately the same complexity, which is less than half of Alg. 1 with the damping factor.
From the above statistical modeling, our target function of marginalization becomes a sparse Bayesian learning (SBL) problem with parameterized Gaussian priors as

\[
f(x, z, \lambda, \alpha) = p(y|z; \lambda) \cdot p(\lambda; c, d) \cdot p(x; \alpha) \cdot \prod_{m=1}^{M} p(\alpha_m; \epsilon, \eta) \prod_{n=1}^{N} \delta(z_n - h_n x).
\] (44)

The statistical modeling parameters \{\lambda, \alpha\} in addition to \{x, z\} are included into the space of variational Bayesian inference. As the parameters \{\epsilon, \eta, c, d\} are considered to be pre-selected, they are not taken as the arguments of \(f(\cdot)\).

Following the standard way, the Bethe free energy can be written as a function of a set of densities, i.e., \(b_{z,\lambda}(z, \lambda), b_{\lambda}(\lambda), b_{x,\alpha}(x, \alpha), \{b_{\alpha_m}(\alpha_m)\}, b_{x,z,n}(x, z_n)\) plus \(b_{x,m}(x_m), b_{z,n}(z_n)\).

For the sake of problem tractability, the constraints on these densities are designed to be

\[
\begin{align*}
  b_{z,\lambda}(z, \lambda) &= b_z(z)b_\lambda(\lambda); \quad (a) \\
  b_{x,\lambda}(x, \alpha) &= b_x(x) \prod_{m=1}^{M} b_{\alpha_m}(\alpha_m); \quad (b) \\
  b_{\alpha_m}(\alpha_m) &= \delta(\alpha_m - \hat{\alpha}_m) \quad (c)
\end{align*}
\] (45)

together with the mean and variance consistency constraints (40) and (42) for the factor densities \(\{b_z(z), b_x(x)\}\) in relation to \(\{b_{x,m}(x_m), b_{z,n}(z_n)\}\). In particular, the factorization constraints in (45)-(a) and -(b) follow the idea for VMP in (18). By doing so we can decouple the correlation between the model parameters \{\alpha, \lambda\} and the latent variables \{x_m, z_n\}. The third one, additionally letting the factor density of \(\alpha_m\) be a Dirac delta-function, reduces VMP to expectation maximization (EM) \[41\]. Such a single-parameter model reduces the complexity for estimating the model parameter \(\alpha_m\) at the cost of accuracy.

Applying the method of Lagrangian multipliers to minimize the Bethe free energy under the above-formulated constraints, we reach to Alg. 4 that combines EM, VMP and Alg. 2 (as the EP variant). In particular, the initialization \(c = d = 0\) ensures a non-informative prior for \(\lambda\). The pair \((\epsilon, \eta)\) on the other hand reflects the sparsity-inducing property of the prior model on \(x\) \[40\]. In particular, \(\epsilon\) plays a dominant role. A smaller \(\epsilon\) encourages a more sparse estimate. Without any prior knowledge of the sparsity of \(x\), we empirically propose in Alg. 4 to adaptively reduce \(\epsilon\) over iterations. The rationale behind the proposal follows the general idea of compressed sensing. Namely, we aim at approximating \(y\) by \(H\hat{x}\) where the support of the estimate \(\hat{x}\) shall be as small as possible.
Knowing $p(x)$ and $\lambda$, Alg. 2 (derived EP variant)
Unknown $p(x)$ and $\lambda$, Alg. 4 (hybrid EM-VMP-Alg.2)
Unknown $p(x)$ and $\lambda$, Fast RVM
Unknown $p(x)$ and $\lambda$, Fast Variational SBL ($\epsilon = \eta = 0$)
Unknown $p(x)$ and $\lambda$, CoSaMP (assume sparsity level $\lfloor \frac{N}{4} \rfloor - 1$)

Figure 4: Normalized mean square error (NMSE) performance versus the sparsity ratio $\rho$ and the number of iterations. This figure is based on the same system configuration as Fig. 3.

Fig. 4 (a) shows that Alg. 4 can approach the performance of Alg. 2 with exact knowledge of $\lambda$ and $p(x)$ when $x$ exhibits low and medium levels of sparsity, i.e., $\rho < 0.3$. With the sparsity beyond 0.4, the measurement ratio $\frac{N}{M} = 0.5$ is too low to yield reliable estimate $\hat{x}$. For comparison, we also include the results from the fast implementation of the relevance vector machine (RVM) [42] and variational SBL (i.e., VMP) [40] together with a non-Bayesian approach CoSaMP [43], [44]. Meanwhile, it is noticed that the algorithms under this investigation are all of similar complexities per iteration. i.e., in an order of $O(MN)$. Further considering their convergence performance in Fig. 4 (b), the additional estimates on hyper parameters would require more iterations. Among algorithms without genie knowledge on $p(x)$ and $\lambda$, the proposed message passing algorithm with a hybrid architecture, i.e., Alg. 4 would potentially of the fastest convergence, while its performance is significantly better.

C. A new viewpoint of message passing algorithms in many wireless applications

Apart from the toy example SSR presented in above, our framework has a wide application in designing communications systems. In this part, we investigate several practical problems in the literature for large-scale wireless communication. They either already used our framework or their ad-hoc solutions fall into our framework. First, the joint channel estimation-equalization-decoding problem addressed in [36] can be equivalently solved by our framework under the
factorization and marginalization consistency constraints. In addition to their hybrid BP-VMP solution, our framework permits to exploit a broad classes of message passing, e.g., EP and AMP for such tasks. Second, our work in [45] has explicitly showcased how to use our framework for robust massive MIMO data detection. In the following, we provide two more examples with a detailed formulation of the constrained Bethe free energy minimization problem.

In [46], an AMP based precoder with low peak-to-average-power-ratio is proposed for massive multi-user MIMO systems. Using the same notations of the original work, we denote the transmit symbol vector, the noise-free received symbol, and the channel coefficients of the \( \mu \)-th user as \( \mathbf{x} \in \mathbb{C}^{N \times 1} \), \( \mathbf{u} \in \mathbb{C}^{M \times 1} \), and \( \mathbf{h}_\mu \in \mathbb{C}^{1 \times N} \), respectively. After introducing the auxiliary vector \( \mathbf{z}_\mu \triangleq \mathbf{h}_\mu \mathbf{x} \), the target marginal \( p(x_i|u) \) is proportional to the outcome of marginalizing

\[
f(\mathbf{z}, \mathbf{x}) = \prod_{\mu=1}^{M} \delta(u_\mu - z_\mu) \prod_{\mu=1}^{M} \delta(z_\mu - \mathbf{h}_\mu \mathbf{x}) \prod_{i=1}^{N} p(x_i),
\]

with respect to \( x_i \). After introducing beliefs \( \{b_{z,\mu}(z_\mu)\}, \{b_{x,\mu}(x)\}, \{b_{x,i}(x_i)\}, \{q_{z,\mu}(z_\mu)\}, \)

one can write the Bethe free energy as

\[
F_B(b) = \sum_{\mu=1}^{M} D[b_{z,\mu}(z_\mu) \| \delta(u_\mu - z_\mu)] + \sum_{i=1}^{N} D[b_{x,i}(x_i) \| p(x_i)]
\]

\[
+ \sum_{i=1}^{N} M H[q_{x,i}(x_i)] + \sum_{\mu=1}^{M} M H[q_{z,\mu}(z_\mu)] - \sum_{\mu=1}^{M} H[b_{x,\mu}(x)].
\]

Considering a constraint set

\[
\forall i \forall \mu \ E[x_i|b_{x,i}] = E[x_i|b_{x,\mu}] = E[x_i|q_{x,i}];
\]

\[
\forall \mu \quad E[z_\mu|b_{z,\mu}] = E[\mathbf{h}_\mu \mathbf{x}|b_{x,\mu}] = E[z_\mu|q_{z,\mu}];
\]

\[
\forall i \quad \text{Var}[x_i|b_{x,i}] = \frac{1}{M} \sum_{\mu=1}^{M} \text{Var}[x_i|b_{x,\mu}] = \text{Var}[x_i|q_{x,i}];
\]

\[
\forall \mu \quad \text{Var}[z_\mu|b_{z,\mu}] = \text{Var}[\mathbf{h}_\mu \mathbf{x}|b_{x,\mu}] = \text{Var}[z_\mu|q_{z,\mu}],
\]

the same algorithm as the original work can be obtained via solving the constrained Bethe free energy minimization problem. Here, we note that, for deriving (G)AMP, the variance consistent constraints in (49) averages out the variance of marginalized belief for \( x_i \), i.e.,

\[
\frac{1}{M} \sum_{\mu=1}^{M} \text{Var}[x_i|b_{x,\mu}] = \text{Var}[x_i|q_{x,i}],
\]

while an EP variant version, as in (42), would require higher complexity with \( \forall \mu, \text{Var}[x_i|b_{x,\mu}] = \text{Var}[x_i|q_{x,i}] \).

Authors of [47] proposed a message passing algorithm termed EM-(G)AMP-Laplacian for
massive MIMO millimeter-wave channel estimation, where the prior distribution of angular-domain channel responses is Laplacian. Using the same notations as in [47], we denote the vectorized angular-domain channel responses, the known coupling matrix formed by the training pilot, the received signal in the angular domain, and the noise vector as $x \in \mathbb{R}^N$, $A \in \mathbb{R}^{M \times N}$, $y \in \mathbb{R}^M$, and $w \in \mathbb{R}^M$, respectively. Under the linear model $y = Ax + w$ and introducing the auxiliary vector $z : z_m \triangleq a_m x$, the target marginal $p(x_n|y)$ is proportional to the outcome of marginalizing

$$f(z, x; \beta) = \prod_{m=1}^M \mathcal{N}(y_m; z_m, \sigma_w^2) \prod_{m=1}^M \delta(z_m - a_m x) \prod_{n=1}^N p(x_n; \beta),$$

(50)

with respect to $x_n$. Here, $a_m$ is the $m$-th row vector of $A$, and the $p(x_n; \beta) = \text{La}(x_n; \beta)$ is a zero-mean Laplacian distribution with scale parameter $\beta$, i.e., $\text{La}(x_n; \beta) = e^{-|x_n|/\beta}/(2\beta)$.

After introducing beliefs $\{b_{z,m}(z_m)\}$, $\{b_{x,m}(x)\}$, $\{b_{x,\beta,n}(x_n, \beta)\}$, $b_{\beta}(\beta)$, $\{q_{x,n}(x_n)\}$, $\{q_{z,m}(z_m)\}$ and factorizing $b_{x,\beta,n}(x_n, \beta)$ as $b_{x,\beta,n}(x_n, \beta) = b_{x,n}(x_n)b_{\beta}(\beta)$ followed by limiting $b_{\beta}(\beta)$ to be $\delta(\beta - \hat{\beta})$, one can write the Bethe free energy as

$$F_B(b) = \sum_{m=1}^M D[b_{z,m}(z_m) \| \mathcal{N}(y_m; z_m, \sigma_w^2)] - \sum_{m=1}^M H[b_{x,m}(x)] + \sum_{n=1}^N D[b_{x,n}(x_n)b_{\beta}(\beta) \| p(x_n; \beta)]$$

$$+ \sum_{n=1}^M M H[q_{x,n}(x_n)] + \sum_{m=1}^M M H[q_{z,m}(z_m)].$$

(51)

The constraint set for the Bethe free energy $F_B(b)$ in above is analogously formulated as the previous example (48) and (49). Solving the constrained Bethe free energy minimization problem by the method of Lagrange multipliers, the same algorithm as in [47] can be obtained.

V. CONCLUSION

This paper unified the message passing algorithms termed BP, EP and VMP under the optimization framework of constrained Bethe free energy minimization. With the same objective function, the key difference of these algorithms simply lies in the way of formulating the constraints. With this identification, it becomes natural to apply proper constraint manipulation for systematically deriving message passing variants from the same framework. In other words, one can treat constraint manipulation as a design freedom, enabling trade-offs between tractability and fidelity of approximate inference.

In particular, we shown that a structured rather than an ad-hoc combination of BP, EP and
VMP can be obtained under a set of partial factorization, marginalization consistency and moment matching constraints. Taking a classic SSR problem as an example, we subsequently derived an EP variant for SSR through constraint reformulation. It is able to outperform the standard EP algorithm with lower computational complexity. Last but not least, hybrid message passing was derived and applied for SBL when the statistical model of the system is unknown. It can deliver the performance that is close to the one attained with perfect knowledge of the statistical model. It can also outperform state-of-the-art solutions in the literature.

In short, constrained Bethe free energy minimization serves as a theoretical framework to perform joint investigation on different message passing algorithms for an improved performance. We focused on constraint manipulation in this work, illustrating its impact on the performance and complexity of message passing. For future works under this framework, it would be interesting to examine other design freedoms, such as the non-convex objective function Bethe free energy. Its construction can be influenced by, but not limited to, factorization of the target function \( f(x) \), design of auxiliary variables (e.g., \( z \) in the examined SSR application), and inclusion of a temperature parameter. It is worth to note that all theses design freedoms are mutually orthogonal, implying the possibility of exploiting them in a combined manner.

**APPENDIX A**

**DERIVATION OF HYBRID VMP-BP**

Firstly, we can get rid of the constraint (18) by interpreting it as a variable interchange, namely substituting \( b_a(x_a) \) with \( \prod_v b_{a,v}(x_{a,v}) \) in the objective function \( F_B(\{b_a\}, \{b_i\}) \) and the other constraint (8). This yields

\[
F_B(\{b_{a,v}\}, \{b_i\}) = \sum_a \int \prod_v b_{a,v}(x_{a,v}) \ln \frac{\prod_v b_{a,v}(x_{a,v})}{f_a(x_a)} dx_a - \sum_i (A_i - 1) \int b_i(x_i) \ln b_i(x_i) dx_i, \quad (A.1)
\]

while the marginalization constraint for each variable node \( x_i \) in (8) is simplified with the mean-field approximation \( b_a(x_a) = \prod_v b_{a,v}(x_{a,v}) \), i.e.,

\[
b_{a,v(i)}(x_i) = b_i(x_i) \quad \forall i \forall a \in A_i
\]

(A.2)

with \( v(i) \) giving \( i \in \mathcal{I}_{a,v(i)} \).

Secondly, we take the method of Lagrange multipliers to solve the problem

\[
\min_{\{b_{a,v}, \{b_i\}\}} F_B(\{b_{a,v}\}, \{b_i\}) \quad \text{s.t.} \quad (A.2).
\]

(A.3)
The Lagrange function is written as

\[ L_B = F_B(\{b_{a,v}\}, \{b_i\}) + \sum_{(a,v)} \zeta_{a,v} \left[ \int b_{a,v}(x_{a,v}) dx_{a,v} - 1 \right] + \sum_i \zeta_i \left[ \int b_i(x_i) dx_i - 1 \right] \]

\[ + \sum_i \sum_{a \in A_i} \sum_{x_i} \lambda_{i \rightarrow a}(x_i) \left[ b_i(x_i) - b_{a,v(i)}(x_i) \right]. \quad \text{(A.4)} \]

In above, we introduce the Lagrange multipliers \( \{\zeta_i, \zeta_{a,v}\} \) for the implicit normalization constraints on the densities \( \{b_{a,v}, b_i\} \), while the additional marginalization consistency constraint (A.2) is associated to the Lagrange multipliers \( \{\lambda_{i \rightarrow a}(x_i)\} \). Here we omit the non-negative constraints on \( \{b_{a,v}, b_i\} \) as later we will find that they are inherently satisfied by any interior stationary point of the Lagrange function.

Thirdly, let us now take the first-order derivatives of \( L_B \) with respect to \( \{b_{a,v}, b_i\} \) and \( \{\lambda_{i \rightarrow a}(x_i)\}, \zeta_i, \zeta_{a,v}\) equal to zeros. By solving the equations, we can express \( \{b_{a,v}, b_i\} \) as

\[ b_{a,v}(x_{a,v}) = e^{-1-\zeta_{a,v}} \prod_{i \in I_{a,v}} e^{\lambda_{i \rightarrow a}(x_i)} \cdot e^\int \Pi_{i' \neq v} b_{a,v'}(x_{a,v'}) \ln f_{a}(x_{a}) dx_{a \setminus v} \]

\[ b_i(x_i) = e^{\zeta_i} \prod_{a \in A_i} \lambda_{i \rightarrow a}(x_i), \quad \text{(A.5)} \]

where the Lagrange multipliers must be a solution of the following equations

\[ e^{\zeta_{a,v} + 1} = \int dx_{a,v} \prod_{i \in I_{a,v}} e^{\lambda_{i \rightarrow a}(x_i)} \cdot \left[ e^\int \Pi_{i' \neq v} b_{a,v'}(x_{a,v'}) \ln f_{a}(x_{a}) dx_{a \setminus v} \right], \quad \text{(A.7)} \]

\[ e^{-\frac{\zeta_i}{A_i-1} + 1} = \int e^{-\frac{1}{A_i-1} \sum_{a \in A_i} \lambda_{i \rightarrow a}(x_i)} dx_i, \quad \text{(A.8)} \]

\[ e^{-\frac{\zeta_{a,v}}{A_i-1} + 1} = e^{-\zeta_{a,v} - \frac{\zeta_i}{A_i-1}} \int dx_{a,v \setminus i} \prod_{i' \in I_{a,v}} e^{\lambda_{i' \rightarrow a}(x_i)} \cdot \left[ e^\int \Pi_{i' \neq v} b_{a,v'}(x_{a,v'}) \ln f_{a}(x_{a}) dx_{a \setminus v} \right]. \quad \text{(A.9)} \]

As we can observe, \( \{b_{a,v}, b_i\} \) in the form of (A.5) and (A.6) are non-negative functions.

Fourthly, attempting to solve the equations of the Lagrange multipliers, the key is to determine \( \{\lambda_{i \rightarrow a}(x_i)\} \) from (A.9), while \( \{\zeta_i, \zeta_{a,v}\} \) can be readily determined by them according to (A.7) and (A.8). For notation simplification, we introduce several auxiliary variables as

\[ \lambda_{a \rightarrow i}(x_i) = \frac{1}{A_i - 1} \sum_{a' \in A_i} \lambda_{i \rightarrow a'}(x_i) - \lambda_{i \rightarrow a}(x_i) \quad \text{(A.10)} \]

\[ m_{a \rightarrow i}(x_i) = e^{\lambda_{a \rightarrow i}(x_i)}, \quad n_{i \rightarrow a}(x_i) = e^{\lambda_{i \rightarrow a}(x_i)}. \quad \text{(A.11)} \]
Using them, the equation (A.9) can be alternatively written as

\[
m_{a\rightarrow i}(x_i) \propto \int dx_{a,v} \prod_{i' \in I_{a,v}} n_{i'\rightarrow a}(x_{i'}) \cdot \left[ e^{\int \prod_{i' \neq v} b_{a,v'}(x_{a,v'}) \ln f_a(x_a) dx_{a,v}} \right] \tag{A.12}
\]

\[
n_{i\rightarrow a}(x_i) = \prod_{a' \in A_i \setminus a} m_{a' \rightarrow i}(x_i). \tag{A.13}
\]

Since \( \{I_{a,v}\} \) are mutually disjoint, the variable \( v \) is actually \( v(i) \) giving \( i \in I_{a,v(i)} \). The solutions to (A.12) and (A.13) yield

\[
b_{a,v}(x_{a,v}) \propto \prod_{i \in I_{a,v}} n_{i\rightarrow a}(x_i) \cdot e^{\int \prod_{i' \neq v} b_{a,v'}(x_{a,v'}) \ln f_a(x_a) dx_{a,v}} \tag{A.14}
\]

\[
b_i(x_i) \propto \prod_{a \in A_i} m_{a\rightarrow i}(x_i), \quad \Rightarrow \quad b_i(x_i) \propto n_{i\rightarrow a}(x_i)m_{a\rightarrow i}(x_i) \quad \forall a \in A_i. \tag{A.15}
\]

After normalization, they correspond to local optima of the constrained Bethe free energy.

**APPENDIX B**

**DERIVATION OF VMP-BP-EP**

Analogously, we follow the method of Lagrange multipliers to solve the problem, starting from constructing the Lagrange function as

\[
L_B = F_B(\{b_{a,v}\}, \{b_i\}) + \sum_{i \in I^{[B]}} \sum_{a \in A_i} \sum_{x_i} \lambda_{i\rightarrow a}(x_i) [b_i(x_i) - b_{a,v(i)}(x_i)] + \sum_i \zeta_i \left[ \int b_i(x_i) dx_i - 1 \right] \\
+ \sum_{(a,v)} \zeta_{a,v} \left[ \int b_{a,v}(x_{a,v}) dx_{a,v} - 1 \right] + \sum_{i \in I^{[B]}} \sum_{a \in A_i} \gamma_{i\rightarrow a}^T [E_{b_i}(t_{a,i}(x_i)) - E_{b_{a,v}(i)}(t_{a,i}(x_i))]. \tag{B.1}
\]

In addition to \( \{\zeta_{a,v}, \zeta_i, \lambda_{i\rightarrow a}(x_i)\} \), we associate the moment matching constraints with the Lagrange multipliers \( \{\gamma_{i\rightarrow a}\} \). The dimension of each vector \( \gamma_{i\rightarrow a} \) is identical to that of the corresponding sufficient statistic \( t_{a,i}(x_i) \).

Let us subsequently set the first-order derivatives of the Lagrange function with respect to the densities to zeros. In doing so, we obtain the density expressions that are in terms of the Lagrange multipliers, namely

\[
b_{a,v}(x_{a,v}) = e^{-1-\zeta_{a,v}} \prod_{i \in I_{a,v}^{[B]}} e^{\lambda_{i\rightarrow a}(x_i)} \prod_{i' \in I_{a,v}^{[B]}} e^{\gamma_{i'\rightarrow a}^T t_{a,i}(x_i)} \cdot e^{\int \prod_{i' \neq v} b_{a,v'}(x_{a,v'}) \ln f_a(x_a) dx_{a,v}} \tag{B.2}
\]

\[
b_i(x_i) = e^{\frac{\zeta_i}{a_i-1}+\frac{1}{a_i-1} \sum_{a \in A_i} \lambda_{i\rightarrow a}(x_i)}, \quad \forall i \in I^{[B]} \tag{B.3}
\]

\[
b_i(x_i) = e^{\frac{\zeta_i}{a_i-1}+\frac{1}{a_i-1} \sum_{a \in A_i} \gamma_{i\rightarrow a}^T t_{a,i}(x_i)}, \quad \forall i \in I^{[E]}, \tag{B.4}
\]
with $\mathcal{I}_{a,v}^{[E]} = \mathcal{I}_{a,v} \cap \mathcal{I}^{[E]}$ and $\mathcal{I}_{a,v}^{[B]} = \mathcal{I}_{a,v} \cap \mathcal{I}^{[B]}$. In (B.4), it is noted that $b_i(x_i)$ is a member of the exponential family $\mathcal{Q}_{a_i}$ that is characterized by the sufficient statistic $t_{a,i}(x_i)$.

By also letting the first-order derivatives of $L_{Bi}$ with respect to the Lagrange multipliers be zeros, the Lagrange multipliers are constrained to ensure that: 1) the above-expressed densities are normalized to one and 2) they fulfill the constraints (24) and (23). In addition to the variable interchanges introduced in (A.10), here we include the following three

$$
\gamma_{a \rightarrow i} = \frac{1}{A_i - 1} \sum_{a' \in A_i} \gamma_{i \rightarrow a'} - \gamma_{i \rightarrow a},
$$

$$
m_{a \rightarrow i}(x_i) = \begin{cases} e^{\lambda_{a \rightarrow i}(x_i)} & i \in \mathcal{I}^{[B]} \\ e^{\gamma_{a \rightarrow i}^T t_{a,i}(x_i)} & i \in \mathcal{I}^{[E]} \end{cases}, \quad n_{i \rightarrow a}(x_i) = \begin{cases} e^{\lambda_{i \rightarrow a}(x_i)} & i \in \mathcal{I}^{[B]} \\ e^{\gamma_{a \rightarrow i}^T t_{a,i}(x_i)} & i \in \mathcal{I}^{[E]} \end{cases}.
$$

Using them, we can establish the fixed-point equations of the Lagrange multipliers as given in (26), (27), and (28).

**ACKNOWLEDGMENT**

We thank Dr. Li You from Southeast University, China and Prof. Bernard H. Fleury from Aalborg University, Denmark, for the insightful discussion that greatly assisted this work.

**REFERENCES**

[1] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[2] D. Barber, *Bayesian Reasoning and Machine Learning*. New York, NY, USA: Cambridge University Press, 2012.

[3] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models,” Mach. Learn., vol. 37, no. 2, pp. 183–233, Nov. 1999.

[4] J. Yedidia, W. Freeman, and Y. Weiss, “Constructing free-energy approximations and generalized belief propagation algorithms,” IEEE Trans. Inf. Theory, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.

[5] J. Winn, C. Bishop, and T. Jaakkola, “Variational message passing,” J. of Machine Learning Research, vol. 6, pp. 661–694, Apr. 2005.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” J. the Royal Statistical Society. Series B, vol. 39, no. 1, pp. 1–38, 1977.

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[8] T. P. Minka, “Expectation propagation for approximate Bayesian inference,” in *Proc. of Conf. on Uncertainty in Artificial Intelligence*, San Francisco, CA, USA, 2001, pp. 362–369.

[9] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, “Gaussian message passing iterative detection for MIMO-NOMA systems with massive access,” in *Proc. IEEE Global Commun. Conf.*, Washington, D.C., USA, Dec. 2016, pp. 1–6.
[10] L. Yang, Y. Liu, and Y. Siu, “Low complexity message passing algorithm for SCMA system,” *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2466–2469, Dec. 2016.

[11] D. Zhang, M. Matté, L. L. Mendes, and G. Fettweis, “Message passing algorithms for upper and lower bounding the coded modulation capacity in a large-scale linear system,” *IEEE Sig. Process. Lett.*, vol. 23, no. 4, pp. 537–540, Apr. 2016.

[12] D. Zhang, L. L. Mendes, M. Matté, I. S. Gaspar, N. Michailow, and G. P. Fettweis, “Expectation propagation for near-optimum detection of MIMO-GFDM signals,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1045–1062, Feb. 2016.

[13] T. L. Hansen, P. B. Jørgensen, M. Badiu, and B. H. Fleury, “An iterative receiver for ofdm with sparsity-based parametric channel estimation,” *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5454–5469, Oct. 2018.

[14] N. J. Myers and R. W. Heath, “Message passing-based joint cfo and channel estimation in mmwave systems with one-bit adcs,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 3064–3077, June 2019.

[15] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” *Proc. of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, Oct. 2009.

[16] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” in *Proc. IEEE Inf. Symp. Inf. Theory*, Saint Petersburg, Russia, Jul. 2011, pp. 2168–2172.

[17] B. Çakmak, O. Winther, and B. H. Fleury, “S-AMP: Approximate message passing for general matrix ensembles,” in *Proc. IEEE Inf. Theory Workshop*, Hobart, Tasmania, Australia, Nov. 2014, pp. 192–196.

[18] X. Meng, S. Wu, L. Kuang, and J. Lu, “An expectation propagation perspective on approximate message passing,” *IEEE Signal Processing Lett.*, vol. 22, no. 8, pp. 1194–1197, Aug. 2015.

[19] S. Wu et al., “Low-complexity iterative detection for estimation with random linear mixing,” *IEEE J. Sel. Areas Signal Process.*, vol. 8, no. 5, pp. 902–915, Oct. 2014.

[20] P. Maechler et al., “VLSI design of approximate message passing for signal restoration and compressive sensing,” *IEEE J. on Emerging and Sel. Topics in Circuits and Syst.*, vol. 2, no. 3, pp. 579–590, Sep. 2012.

[21] S. Wang et al., “Energy-efficient and low-complexity uplink transceiver for massive spatial modulation MIMO,” *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4617–4632, Oct. 2015.

[22] Y. Zhang, L. Huang, and J. Song, “MPDQ-based high-order QAM detection scheme for massive MIMO with low-precision quantization,” in *Proc. IEEE Global Conf. on Signal and Inf. Process.*, Orlando, Florida, USA, Dec. 2015, pp. 28–32.

[23] S. Lyu and C. Ling, “Fast variational sparse Bayesian learning with automatic relevance determination for superimposed signals,” in *Proc. IEEE Int. Conf. on Ubiquitous Wireless Broadband*, Paris, France, Oct. 2016, pp. 1–4.

[24] E. Karseras et al., “Computationally efficient sparse Bayesian learning via generalized approximate message passing,” in *Proc. IEEE Int. Conf. on Ubiquitous Wireless Broadband*, Paris, France, Oct. 2016, pp. 1–4.

[25] D. Shutin, T. Buchgraber, S. Kulkarni, and H. Poor, “Fast variational sparse Bayesian learning with automatic relevance determination for superimposed signals,” *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 6257–6261, Dec. 2011.

[26] X. Zou et al., “Fast variational Bayesian learning for channel estimation with prior statistical information,” in *Proc. IEEE Int. Workshop on Signal Process. Adv. in Wireless Commun.*, Stockholm, Sweden, Jun. 2015, pp. 470–474.

[27] J. Mo et al., “Channel estimation in millimeter wave MIMO systems with one-bit quantization,” in *Asilomar Conf. on Signals, Syst. and Comput.*, Pacific Grove, CA, USA, Nov. 2014, pp. 957–961.

[28] C. Stöckle et al., “Channel estimation in massive MIMO systems using 1-bit quantization,” in *Proc. IEEE Int. Workshop on Signal Process. Adv. in Wireless Commun.*, Jul. 2016, pp. 1–6.
[30] Z. Utkovski et al., “Random access in C-RAN for user activity detection with limited-capacity fronthaul,” IEEE Signal Process. Lett., vol. 24, no. 1, pp. 17–21, Jan. 2017.

[31] M. Senst and G. Ascheid, “A combined belief propagation and mean field algorithm for soft carrier phase estimation,” in Proc. IEEE Inter. Symp. on Wireless Commun. Syst., Aachen, Germany, Nov. 2011, pp. 512–516.

[32] M. A. Badiu, G. E. Kirkelund, C. N. Manchón, E. Riegler, and B. H. Fleury, “Message-passing algorithms for channel estimation and decoding using approximate inference,” in Proc. IEEE Inter. Symp. on Inf. Theory, MA, USA, Jul. 2012, pp. 2376–2380.

[33] P. Sun, C. Zhang, Z. Wang, C. Manchon, and B. Fleury, “Iterative receiver design for ISI channels using combined belief- and expectation-propagation,” IEEE Signal Processing Lett., vol. 22, no. 10, pp. 1733–1737, Oct. 2015.

[34] W. Wang, Z. Wang, C. Zhang, Q. Guo, P. Sun, and X. Wang, “A BP-MF-EP based iterative receiver for joint phase noise estimation, equalization, and decoding,” IEEE Signal Processing Lett., vol. 23, no. 10, pp. 1349–1353, Oct. 2016.

[35] N. Wu, W. Yuan, Q. Guo, and J. Kuang, “A hybrid BP-EP-VMP approach to joint channel estimation and decoding for FTN signaling over frequency selective fading channels,” IEEE Access, vol. 5, pp. 6849–6858, May 2017.

[36] E. Riegler, G. Kirkelund, C. Manchon, M. Badiu, and B. Fleury, “Merging belief propagation and the mean field approximation: A free energy approach,” IEEE Trans. Inf. Theory, vol. 59, no. 1, pp. 588–602, Jan. 2013.

[37] F. R. Kschischang, B. J. Frey, and H. A. Loeliger, “Factor graphs and the sum-product algorithm,” IEEE Trans. Inf. Theory, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[38] T. Heskes, “Stable fixed points of loopy belief propagation are minima of the Bethe free energy,” in Proc. Neural Inf. Process. Syst., Nevada, USA, 2003.

[39] T. Heskes and O. Zoeter, “Expectation propagation for approximate inference in dynamic Bayesian networks,” in Proc. Conf. on Uncertain. in Artif. Intell., Alberta, Canada, 2002, pp. 216–223.

[40] N. L. Pedersen, C. N. Manchón, M.-A. Badiu, D. Shutin, and B. H. Fleury, “Sparse estimation using Bayesian hierarchical prior modeling for real and complex linear models,” Signal Processing, vol. 115, pp. 94–109, Mar. 2015.

[41] J. Dauwels, “On variational message passing on factor graphs,” in Proc. IEEE Inf. Symp. Inf. Theory, Nice, France, Jun. 2007, pp. 2546–2550.

[42] M. E. Tipping and A. Faul, “Fast marginal likelihood maximisation for sparse bayesian models,” in Proc. Int. Workshop on Artificial Intelligence and Statistics, Florida, USA, Jan. 2003, pp. 3–6.

[43] D. Needell and J. A. Tropp, “CoSaMP: Iterative signal recovery from incomplete and inaccurate samples,” Commun. ACM, vol. 53, no. 12, pp. 93–100, Dec. 2010.

[44] S. Becker, Matlab code: CoSaMP and OMP for sparse recovery, Aug. 2016. [Online]. Available: https://de.mathworks.com/matlabcentral/fileexchange/32402-cosamp-and-omp-for-sparse-recovery

[45] S. Chen, W. Wang, D. Zhang, and X. Gao, “Robust approximate message passing detection based on minimizing bethe free energy for massive mimo systems,” in IEEE Vehicular Technology Conference, Sep. 2017, pp. 1–6.

[46] J. Chen, C. Wang, K. Wong, and C. Wen, “Low-complexity precoding design for massive multiuser mimo systems using approximate message passing,” IEEE Transactions on Vehicular Technology, vol. 65, no. 7, pp. 5707–5714, July 2016.

[47] F. Bellili, F. Sohrabi, and W. Yu, “Generalized approximate message passing for massive mimo mmwave channel estimation with laplacian prior,” IEEE Transactions on Communications, vol. 67, no. 5, pp. 3205–3219, May 2019.