EFFICIENT BAYESIAN INFERENCE BY CONJUGATE-COMPUTATION VARIATIONAL MESSAGE PASSING

Mykola Lukashchuk[1,*] *İsmail Şenöz*[1,*] *Bert de Vries*[1,2]

[1]Department of Electrical Engineering, Eindhoven University of Technology
[2]GN Hearing, Eindhoven, the Netherlands
{m.lukashchuk, i.senoz, bert.de.vries}@tue.nl
[*]These authors contributed equally to this work

ABSTRACT

Variational message passing is an efficient Bayesian inference method in factorized probabilistic models composed of conjugate factors from the exponential family (EF) distributions. In many applications, a more accurate model for the process under consideration can be obtained by inserting nonlinear deterministic factors in the model. Unfortunately, variational messages that pass through nonlinear nodes cannot be analytically computed in closed form. In this paper, we derive an efficient algorithm for passing variational messages through arbitrary deterministic factors. Our method is based on projecting outgoing messages onto an EF distribution. We implemented our algorithm in RxInfer, which is an open-source message passing-based inference package in Julia. The resulting implementation yields efficient message passing-based inference in arbitrary models composed of stochastic and deterministic factors. We compare our method to alternative state-of-the-art inference methods and find lower (i.e., better) free energy residuals for the proposed method.

Index Terms— Factor graphs, Non-linear Filtering, Message Passing, Variational Bayesian inference

1 Introduction

Bayesian reasoning, committed to using probability theory for handling uncertainty, is optimal under the assumption of universally agreeable axioms [1]. This perspective suggests viewing parameter inference in models such as state space and hidden Markov models as an exercise in Bayesian reasoning. However, the complex integral computations in Bayes' rule necessitate algorithms that strike a balance between accuracy and computational efficiency. In line with this, our paper presents an improved algorithm that minimizes computational costs while maintaining competitive accuracy.

We shortly discuss the positioning of the proposed algorithm relative to existing work in the context of Fig. 1. Analytical closed-form solutions for Bayesian inference tasks are only available for linear Gaussian systems. For more complex systems, Bayesian inference is approximated by Monte Carlo

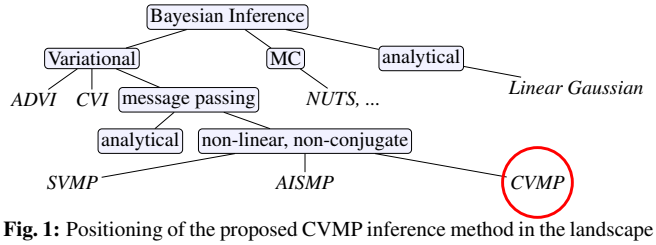

Fig. 1: Positioning of the proposed CVMP inference method in the landscape of Bayesian inference methods. This view on inference methods is not exhaustive nor unique. See discussion in section 1 for interpretation.

(MC) sampling or by variational optimization of a tractable bound on Bayesian evidence. Monte Carlo sampling methods such as the No-U-Turn Sampler (NUTS) [2] are not suited for real-time inference in signal processing tasks with strong resource constraints. A popular variant of variational inference is Automated Differentiation Variational Inference (ADVI) [3], which is very user-friendly (i.e., automated), but is also computationally heavy. Another important algorithm in this space is Kahn's Conjugate-computation Variational Inference (CVI) algorithm that extends natural gradient descent to non-conjugate models [4]. However, for real-time signal processing in a non-trivial model, we will argue in section 2 that message passing-based inference is arguably the only feasible method [5, 6]. In particular, the Variational Message Passing (VMP) algorithm provides closed-form message update rules in models composed of appropriately matched ("conjugate") prior and likelihood pairs from the exponential family of distributions. Unfortunately, message computation in non-conjugate and non-linear models is not analytically solvable, and various approximate solutions have been proposed [7, 8]. Recently, [9] proposed Stochastic Variational Message Passing (SVMP), which casts Kahn's CVI algorithm into a factor graph framework. This is an important development for the signal processing community since fast and accurate inference for a wide range of non-conjugate, non-linear models has become a reality. In the chase for low-complexity high-accuracy inference, SVMP still features a weakness as its forward message update rule uses a list of samples to represent a probability distribution. This is both incompatible with other

979-8-3503-2411-2/23/$31.00 ©2023 European Union

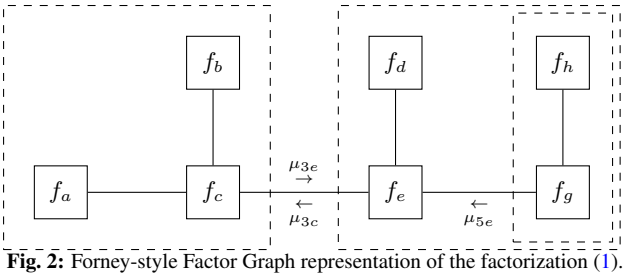

Fig. 2: Forney-style Factor Graph representation of the factorization (1).

(closed-form) messages and potentially computationally expensive. In this paper, we derive a closed-form forward message update rule that combines well with SVMP's backward message and interfaces smoothly with conventional message passing rules. To honor Kahn's CVI approach, we call our proposed algorithm Conjugate-computation Variational Message Passing (CVMP).

2 Background on Factor Graphs and Variational Inference

We shortly recapitulate why message passing in factor graphs is an attractive inference method for signal processing tasks. Consider a factorized multivariate function

$$f(x_1, x_2, \ldots, x_6) = f_a(x_1) f_b(x_2) f_c(x_3, x_2, x_1) \\ f_d(x_4) f_e(x_5, x_4, x_3) f_g(x_6, x_5) f_h(x_6)\,. \tag{1}$$

Assume that we are interested in inferring (the so-called marginal distribution)

$$q(x_3) \triangleq \sum_{x_1, x_2, x_4, x_5, x_6} f(x_1, x_2, \ldots, x_6)\,. \tag{2}$$

If each variable x_i in (2) has about 10 possible values, then the sum contains about 1 million terms. However, making use of the factorization (1) and the distributive law, we can rewrite this sum as $q(x_3) = \mu_{3c}(x_3)\mu_{3e}(x_3)$ where

$$\mu_{3e}(x_3) \triangleq \sum_{x_1, x_2} f_a(x_1) f_b(x_2) f_c(x_3, x_2, x_1) \tag{3a}$$

$$\mu_{3c}(x_3) \triangleq \sum_{x_4, x_5} f_d(x_4) f_e(x_5, x_4, x_3) \mu_{5e}(x_5) \tag{3b}$$

$$\mu_{5e}(x_5) \triangleq \sum_{x_6} f_g(x_6, x_5) f_h(x_6) \tag{3c}$$

The computation in (3), which requires only a few hundred summations and multiplications, is clearly preferred from a computational load viewpoint. To execute (3), we need to compute intermediate results $\mu_{ai}(x_i)$ and $\mu_{ia}(x_i)$ that afford an interpretation of local messages in a Forney-style Factor Graph (FFG) representation of the model, see Fig. 2.

An FFG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, represents a factorized function,

$$f(x) = \prod_{a \in \mathcal{V}} f_a(x_a)\,, \tag{4}$$

where x_a collects the argument variables of factor f_a. We assume that all the factors are non-negative. In an FFG, a node $a \in \mathcal{V}$ corresponds to a factor f_a, and the neighboring edges $\mathcal{E}(a)$ correspond to the variables x_a. An edge is connected to a node if the variable of that edge is an argument of the factor of the node. We denote the neighboring edges of a node $a \in \mathcal{V}$ by $\mathcal{E}(a)$. Vice versa, for an edge $i \in \mathcal{E}$, the notation $\mathcal{V}(i)$ collects all neighboring nodes. As a notational convention, we index nodes by a, b, c and edges by i, j, unless stated otherwise. In this paper, we will frequently refer to a subgraph. We define an edge-induced subgraph by $\mathcal{G}(i) = (\mathcal{V}(i), i)$, and a node-induced subgraph by $\mathcal{G}(a) = (a, \mathcal{E}(a))$. Furthermore, we denote a local subgraph by $\mathcal{G}(a, i) = (\mathcal{V}(i), \mathcal{E}(a))$, which collects all local nodes and edges around i and a respectively. The FFG formalism of Forney constrains $\max|\mathcal{V}(i)| = 2$ for every $i \in \mathcal{E}$, but we adhere to the terminated FFG formalism of [6], $|\mathcal{V}(i)| = 2$ for every $i \in \mathcal{E}$, by terminating half-edges with a factor that is proportional to 1.

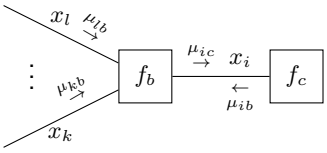

Fig. 3: Visualization of a subgraph with indicated sum-product messages.

Following (3), inference in an FFG is settled by the computation of outgoing messages from nodes. Marginalization results from the product of colliding messages on the edges. The efficiency of Bayesian inference by message passing in a factor graph is therefore *entirely determined by the accuracy and costs of computing messages and the product of colliding messages*. For a subgraph $\mathcal{G}(b, i) = (\mathcal{V}(i), \mathcal{E}(b))$ induced by a factor f_b and variable x_i as displayed in Fig. 3, [6, Theorem 1] derives the following update equations for the outgoing message μ_{ic} from node b and to the node c along the edge i, and the marginal $q_i(x_i)$ for variable x_i:

$$\mu_{ic}(x_i) = \int f_b(x_b) \prod_{\substack{j \in \mathcal{E}(b) \\ j \neq i}} \mu_{jb}(x_j) \mathrm{d}x_j \tag{5}$$

$$q_i(x_i) = \frac{\mu_{ib}(x_i)\mu_{ic}(x_i)}{\int \mu_{ib}(x_i)\mu_{ic}(x_i)\mathrm{d}x_i}\,. \tag{6}$$

Marginal update equations (6) are derived as the stationary solutions of the Bethe free energy (BFE) augmented with marginalization constraints [10, 6] where the messages (5) are obtained as the exponentiated Lagrange multipliers that enforce marginalization constraints. Rigorously, given a subgraph $\mathcal{G}(b, i)$, the local stationary points of the minimization problem $\arg\min_q L[q, f]$, where the Lagrangian $L[q, f]$ is

$$L[q, f] = \sum_{a \in \mathcal{V}} D_{\mathrm{KL}}[q_a || f_a] + \sum_{a \in \mathcal{V}} \psi_a \left[\int q_a(x_a) \mathrm{d}x_a - 1 \right] \\ + \sum_{a \in \mathcal{V}} \sum_{i \in \mathcal{E}(a)} \int \lambda_{ia}(x_i) \left[q_i(x_i) - \int q_a(x_a) \mathrm{d}x_{a \setminus i} \right] \mathrm{d}x_i \\ + \sum_{i \in \mathcal{E}} H[q_i] + \sum_{i \in \mathcal{E}} \psi_i \left[\int q_i(x_i) \mathrm{d}x_i - 1 \right]\,, \tag{7}$$

are given by (6) as the product of messages (5) that ensures marginalization constraints with the definition $\mu_{ia} \triangleq \exp(\lambda_{ia})$ [10, 6].

In general, the literature about inference in factor graphs is about efficient approximations of (5) and (6) for a very wide range of (both stochastic and deterministic, continuous, and discrete) factors [6, 11, 12].

3 Problem Statement

Given an FFG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ representing a factorized function (4), we consider subgraphs of the form $\mathcal{G}(b,i) = (\mathcal{V}(i), \mathcal{E}(b))$ where $f_b(x_b) = \delta(x_i - h_b(x_{b\setminus i}))$ is an implicit function considered as a deterministic factor with x_i being the output of the function $h_b(x_{b\setminus i})$. In the rest of the paper, without loss of generality, we will assume $\mathcal{V}(i) = \{b, c\}$. Figure 3 illustrates the subgraphs of this form. These subgraph structures are ubiquitous in many generative models. Due to non-linearities introduced by deterministic factors, the computation of messages and (joint) marginal distributions for subgraphs $\mathcal{G}(b,i)$ is challenging.

Moreover, when an FFG includes hard constraints, due to [13, Conjecture 1, Conjecture 2], the stationary points of (7) obtained by [10, Theorem 2][6, Theorem 1] are not necessarily interior points (edge minima), which might cause the Lagrange multipliers to have logarithmically-divergent behavior that causes the marginals (6) and messages (5) to be zero. If one accepts to work with zero beliefs, then no problem arises. However, if zero beliefs are unacceptable, one must accept to work with approximations to avoid edge minima of (7).

Equations (5) and (6) are often not available in closed form, and brute force computations are not feasible within reasonable time limits constrained by limited computing power. This paper addresses the problem of approximating the message (5) and marginal computations (6) for deterministic factors by closed-form solutions.

4 Solution Proposal

To address the approximation of (5) and (6), we adhere to the projection recipe derived in [11], which allows us to obtain closed-form approximations to messages and marginals. Our solution proposal relies on the minimization of KL divergence locally for the subgraphs of the form $\mathcal{G}(b,i)$. Our focus will be obtaining the message and marginal for the edge $i \in \mathcal{E}(b)$ that is the output of the nonlinear mapping $h_b(x_{b\setminus i})$.

In the rest of this section, we rigorously derive the new update equations for the forward message through a non-linear deterministic node. The resulting algorithm (CVMP) is displayed in Algorithm 1.

Theorem 1. *Given a subgraph $\mathcal{G}(b,i)$ where $f_b(x_b) = \delta(x_i - h_b(x_{b\setminus i}))$ is a deterministic factor, assume that the messages μ_{jb} for all $j \in \mathcal{E}(b)$ are given. Assume that the message $\mu_{ib}(x_i) \propto \exp\left(\eta_{ib}^\top T_i(x_i)\right)$ is in an exponential family or a projection of it to an exponential family exists. Further, assume that $q_{b\setminus i}(x_{b\setminus i})$ is available. Then, the stationary solutions to the following minimization problem,*

$$\arg\min_{\eta_i \in \Omega_i} D_{KL}[q_{\eta_i} || q_i] \,, \tag{8}$$

where $q_{b\setminus i}, q_i = \arg\min_{q_i, q_{b\setminus i}} L[q, f]$ as obtained by [6, Theorem 1], are in one-to-one correspondence with the fixed points of the following iterations:

$$\eta_i^{(k+1)} = \eta_i^{(k)} + \mathcal{I}^{-1}\left(\eta_i^{(k)}\right) \nabla_{\eta_i} \mathbb{M}\left[q_{\eta_i}^{(k)}(h_b(x_{b\setminus i}))\right] \,, \tag{9}$$

where $q_i(x_i)$ is defined as per (6), k is an iteration index, $\mathcal{I}(\eta_i) = \mathbf{H}_{A_i}(\eta_i)$ is the Fisher information matrix, i.e., the Hessian of log partition $A_i(\eta_i)$ [1], and

$$\mathbb{M}\left[q_{\eta_i}(h_b(x_{b\setminus i}))\right] \triangleq \int \log q_{\eta_i}(h_b(x_{b\setminus i})) |J_{h_b}(x_{b\setminus i})| \mathrm{d}x_{b\setminus i} \tag{10}$$

$$\mathrm{d}x_{b\setminus i} = q_{b\setminus i}(x_{b\setminus i}) \prod_{\substack{j \in \mathcal{E}(b) \\ j \neq i}} \mu_{jb}(x_j) \mathrm{d}x_j \tag{11}$$

$$q_{\eta_i}(x_i) = h_i(x_i) \exp(\eta_i^\top T_i(x_i) - A_i(\eta_i)) \tag{12}$$

Proof. *First, we show that the solutions of $\nabla_{\eta_i} D_{KL}[q_{\eta_i} || q_i] = 0$ satisfy the fixed point iterations (9). Using [9, Appendix A], we can write the gradient of the KL divergence as follows*

$$\nabla_{\eta_i} D_{KL} = \mathcal{I}(\eta_i)\,(\eta_i - \eta_{ib}) - \nabla_{\eta_i} \mathbb{E}_{q_{\eta_i}}\left[\log \mu_{ic}(x_i)\right] . \tag{13}$$

Using factor $f_b(x_b)$, we obtain the following identity

$$q_{b\setminus i}(x_{b\setminus i}) = \int \delta(x_i - h_b(x_{b\setminus i})) q_{\eta_i}(x_i) \mathrm{d}x_i = q_{\eta_i}(h_b(x_{b\setminus i}))$$

This means we can transform the second term in (13) as follows by change of variables $x_i = h_b(x_{b\setminus i})$[2]

$$\begin{aligned}
&\nabla_{\eta_i} \mathbb{E}_{q_{\eta_i}}\left[\log \mu_{ic}\right] \\
&\quad = \nabla_{\eta_i} \int q_{b\setminus i}(x_{b\setminus i}) \log \frac{q_{\eta_i}(h_b(x_{b\setminus i}))}{\mu_{ib}(h_b(x_{b\setminus i}))} |J_{h_b}(x_{b\setminus i})| \mathrm{d}x_{b\setminus i} \\
&\quad = \nabla_{\eta_i} \int q_{b\setminus i}(x_{b\setminus i}) \log q_{\eta_i}(h_b(x_{b\setminus i})) |J_{h_b}(x_{b\setminus i})| \mathrm{d}x_{b\setminus i} \,,
\end{aligned}$$

where the third line follows because μ_{ib} is not included as a multiplying factor involving η_i. Then we can write the differential using the positive valued messages $\mu_{jb}(x_j)$ for $j \neq i$ as the following measure $\mathrm{d}x_{b\setminus i} = \prod_{j \neq i} \mu_{jb}(x_j) \mathrm{d}x_j$. Together with this differential, we identify the last line as $\nabla_{\eta_i} \mathbb{M}\left[q_{\eta_i}(h_b(x_{b\setminus i}))\right]$ defined in (10). Solving, $\mathcal{I}(\eta_i)(\eta_i - \eta_{ib}) = \nabla_{\eta_i} \mathbb{M}\left[q_{\eta_i}(h_b(x_{b\setminus i}))\right]$ we obtain the stationary point

$$\eta_i = \eta_{ib} + \mathcal{I}^{-1}(\eta_i) \nabla_{\eta_i} \mathbb{M}\left[q_{\eta_i}(h_b(x_{b\setminus i}))\right] \,, \tag{14}$$

which satisfies the fixed point iterations (9).

To show the reverse direction, we assume that there exist fixed points of (9). This means that there exists K such that $\eta_i^{(k+1)} = \eta_i^{(k)}$ for all $k > K$. Hence, for all $k > K$, plugging $\eta_i^{(k+1)}$ and $\eta_i^{(k)}$ into (9) we obtain $\mathcal{I}^{-1}\left(\eta_i^{(k)}\right) \nabla_{\eta_i} \mathbb{M}\left[q_{\eta_i}^{(k)}\right] = 0$. Since the Fisher information matrix $\mathcal{I}(\eta_i)$ is positive definite, its inverse is also positive definite. This implies that $\nabla_{\eta_i} \mathbb{M}\left[q_{\eta_i}^{(k)}\right] = 0$ for all $k > K$. What remains is to evaluate the gradient of the KL divergence at $\eta_i^{(k+1)}$. Using (13) we

[1] Because log partition is convex [14, Proposition 3.1], its Hessian is always positive definite. Technically, Fisher information matrix is positive semi-definite. But here, we use Fisher information to refer to strictly positive-definite Hessian of the log partition function.

[2] Change of variable does not influence KL divergence since it is invariant to coordinate transformations.

obtain $\nabla_{\eta_i^{(k+1)}} D_{KL} = 0$ *since*

$$\mathcal{I}\left(\eta_i^{(k+1)}\right)\left(\eta_i^{(k+1)} - \eta_i^{(k)}\right) - \nabla_{\eta_i}\mathbb{M}\left[q_{\eta_i}^{(k+1)}\right] = 0\,.$$

This means that the exponential family distribution whose natural parameter is obtained by the fixed point iterations (9) is a stationary solution to the minimization problem (8). □

Remark. *Theorem 1 allows us to obtain an exponential family approximation to the marginal (6) that is obtained by minimization of the Bethe free energy [10, Theorem 2][6, Theorem 1]. In Theorem 1, we construct a locally convex upper bound to the Bethe free energy for the deterministic factors.*

Lemma 2. *Assume that a fixed point* $\eta_i^\star$ *of the iterations (9) exist, the message* $\mu_{ib}(x_i) \propto \exp(\eta_{ib}^\top T_i(x_i))$ *is available and further assume that* $\mu_{ic}(x_i) \propto \exp\left(\eta_{ic}^\top T_{ic}(x_i)\right)$*. Then the roots of the following expression:*

$$\nabla_{\eta_i} A_i(\eta_i^\star - \eta_{ib}) - \nabla_{\eta_{ic}} \int \mu_{\eta_{ic}}(x_i)\mathrm{d}x_i \qquad (15)$$

are stationary solutions to the following reverse KL minimization problem[3]*:*

$$\arg\min_{\eta_{ic}\in\Omega_i}\left(D_{KL}[\mu_{ic}||\mu_{\eta_{ic}}] + \int(\mu_{\eta_{ic}}(x_i) - \mu_{ic}(x_i))\mathrm{d}x_i\right). \quad (16)$$

Proof. *We take the gradient of (16) and solve*

$$\nabla_{\eta_{ic}}\left(\int \mu_{\eta_{ic}}(x_i)\mathrm{d}x_i - \mathbb{E}_{\mu_{ic}}[\log \mu_{\eta_{ic}}]\right) = 0$$

$$\nabla_{\eta_{ic}}\left(\int \mu_{\eta_{ic}}(x_i)\mathrm{d}x_i - \int \frac{q_i(x_i)}{\mu_{ib}(x_i)}\log \mu_{\eta_{ic}}\mathrm{d}x_i\right) = 0$$

$$\nabla_{\eta_{ic}}\int \mu_{\eta_{ic}}(x_i)\mathrm{d}x_i - \nabla_{\eta_i} A_i(\eta_i^\star - \eta_{ib}) = 0\,,$$

which proves that the roots are the stationary solutions. □

In practice, a root-finding algorithm can be used to compute the roots of (15). However, a simple solution exists if we constrain the form of the message μ_{ic} to be the same as μ_{ib} by assuming the same sufficient statistics, then the roots can be obtained by the following corollary.

Corollary 3. *Given the assumptions of Lemma 2, further assume that* $T_{ic}(x_i) = T_i(x_i)$*. Then the stationary solutions of (16) are given by* $\eta_{ic} = \eta_i^\star - \eta_{ib}$*.*

Proof. *Since* $T_{ic}(x_i) = T_i(x_i)$*, we have*

$$\nabla_{\eta_i} A_i(\eta_i^\star - \eta_{ib}) = \nabla_{\eta_{ic}}\int \mu_{\eta_{ic}}(x_i)\mathrm{d}x_i = \nabla_{\eta_i} A_i(\eta_{ic})\,. \quad (17)$$

But, then the triviality $\eta_{ic} = \eta_i^\star - \eta_{ib}$ *is a root of (15) due to (17), hence a stationary solution of (16) by Lemma 2.* □

Expectations that are required to compute natural gradients (9) are often not easy to compute in closed form. We resort to the popular REINFORCE estimator [15].

Remark. *We compute natural gradient estimates by the REINFORCE estimator using the following approximation obtained via Monte-Carlo summation by using samples* $x_{b\setminus i}^{(s)}$ *from* $q_{\eta_{b\setminus i}}(x_{b\setminus i})$ *and the natural parameter* η_{ib} *of the incoming message* $\mu_{ib}(x_i)$*:*

$$\nabla_{\eta_i}\mathbb{M}\left[q_{\eta_i}\right] \approx \mathcal{I}^{-1}(\eta_i)\left(\frac{1}{S}\sum_s \nabla_{\eta_i}\left(\mu_{\eta_i}\left(x_{b\setminus i}^{(s)}\right)\right)\right), \quad (18)$$

where the natural gradient is defined as

$$\tilde{\nabla}_{\eta_i}\mathbb{M}\left[q_{\eta_i}\right] \triangleq \eta_i - \left(\eta_{ib} + \mathcal{I}^{-1}(\eta_i)\nabla_{\eta_i}\mathbb{M}\left[q_{\eta_i}\right]\right), \quad (19)$$

and, for convenience, we have defined

$$\mu_{\eta_i}\left(x_{b\setminus i}^{(s)}\right) \triangleq \log q_{\eta_i}\left(h_b\left(x_{b\setminus i}^{(s)}\right)\right)\left|J_{h_b}\left(x_{b\setminus i}^{(s)}\right)\right| \prod_{\substack{j\in\mathcal{E}(b)\\ j\neq i}} \mu_{jb}\left(x_j^{(s)}\right).$$

Algorithm 1 Conjugate-computation Variational Message Passing (CVMP)

Input A subgraph $\mathcal{G}(b,i)$ induced by a factor f_b and an edge $i \in \mathcal{E}(b)$ such that $f_b(x_b) = \delta(x_i - h_b(x_{b\setminus i}))$, messages μ_{jb} for all $j \in \mathcal{E}(b)$, the marginal $q_{b\setminus i}$, exponential family message μ_{ib} with natural parameter η_{ib} in accordance with Theorem 1, a Robbins-Monro sequence ρ_k, and tolerance $\epsilon > 0$

Output $\mu_{ic}(x_i)$ and $q_{\eta_i}(x_i)$

procedure CVMP
- Draw samples $x_{b\setminus i}^{(s)} \sim q_{\eta_{b\setminus i}}(x_{b\setminus i})$
- **repeat**
  - Compute natural gradient estimate $\tilde{\nabla}_{\eta_i}^{(k)}$ by (19)
  - Update $\eta_i^{(k+1)} = \eta_i^{(k)} - \rho_k\tilde{\nabla}_{\eta_i}^{(k)}$
- **until** $\left|\left|\rho_k\tilde{\nabla}_i^{(k)}\right|\right| < \varepsilon$
- Update $q_{\eta_i}(x_i) \propto \exp(\eta_i^\top T_i(x_i))$
- Solve (15) (or Corollary 3) to obtain η_{ic}
- Update $\mu_{ic}(x_i) \propto \exp\left(\eta_{ic}^\top T_{ic}(x_i)\right)$

end procedure

5 Experiments

We apply CVMP to analyze annual solar activities from 1945 to 2020, similar to experiments in [16]. The sunspots dataset is sourced from the WDC-SILSO, Royal Observatory of Belgium [17]. The data samples were rounded to their closest integer values. We implemented our algorithm in the RxInfer [18]. Our implementation is readily available.[4]

Our primary objective is to illustrate the improvement offered by CVMP in passing forward messages through nonlinear deterministic nodes when compared to SVMP.

The generative model is specified as

$$p(\gamma) = \Gamma(\gamma|1000, 1), p(z_0|\gamma) = \Gamma(z_0|1, \gamma) \quad (20a)$$

$$p(z_t|z_{t-1}) = \delta(z_t - h(z_{t-1})) \quad (20b)$$

$$p(x_t|z_t, \gamma) = \Gamma(x_t|z_t, \gamma) \quad (20c)$$

$$p(y_t|x_t) = \mathrm{Pois}(y_t|x_t) \quad (20d)$$

where $h(z) \triangleq \log(\exp(z) + 1)$, $\Gamma(\cdot|a, b)$ denotes the Gamma distribution with shape a and rate b, and $\mathrm{Pois}(\cdot|\gamma)$ is the Poisson distribution with rate parameter γ.

The inference goal was to compute the posterior $q(z_{1:T})$. We assumed a mean-field factorization

$$q(z_{1:T}) = \prod_{t=1}^{T} q(z_t). \quad (21)$$

[3] We put the integral to the right-hand side of (16) to account for the fact that messages are not necessarily normalized

[4] https://github.com/biaslab/CVMP

Table 1: Inference results for the sunspot data experiment [17]. We compare results of CVMP, SVMP, AISMP, and NUTS in terms of Bethe Free Energy (BFE), Root Mean Square Error (RMSE), and inference exectution time. Two distinct versions of CVMP are presented: a "fast" version (CVMP^1, with 20 samples drawn from $q_{\eta_{b\setminus i}}(x_{b\setminus i})$) and an "accurate" version (CVMP^2, with 1000 samples).

Method	BFE	RMSE	Inference Time
CVMP^1	5861.64	27.33	0.272s
CVMP^2	5733.71	21.23	6.371s
SVMP	6262.77	34.4	0.247s
AISMP	6696.54	20.84	132.76s
NUTS	*	20.1	14.602s

In terms of performance assessment, we report the minimized Bethe Free Energy (BFE), which is a special case of variational Free Energy [13]. The variational free energy can be decomposed as "complexity of computation" minus "inference accuracy". Hence, BFE minimization by message passing aims for maximal accuracy at minimal (computational) costs. Since NUTS does not provide BFE, we also report the RMSE between the mean of the posterior $q(z_{1:T})$ and the data (the hidden signal) from the sunspot dataset [17] as a measure of accuracy. Additionally, we report the execution time on an Apple M1 Pro Chip with 8 cores, and 32GB RAM.

The hyperparameters in CVMP, i.e., a Robbins-Monro sequence (ρ_k) and a tolerance parameter (ϵ), were selected using the Descent optimizer from the Flux package [19].

After running CVMP for 10 iterations, the Bethe free energy converged. The distributions mean and 95% confidence intervals of the posterior are visualized in Figure 4. We compare the estimates produced by CVMP with those obtained from SVMP, AISMP, and NUTS. Table 1 shows that while NUTS and AISMP provide better estimates in terms of RMSE, CVMP achieves a significant reduction in inference time compared to AISMP, and reaches a lower (i.e., better) BFE and RMSE than SVMP.

The experimental results illustrate CVMP's superior BFE over SVMP and AISMP, while maintaining significantly lower execution time compared to AISMP and NUTS. However, note that AISMP and NUTS achieve better RMSE values than CVMP and therefore may be preferred in scenarios where execution time is not a priority.

CVMP balances AISMP and SVMP, providing an optimal trade-off between BFE, RMSE, and computational efficiency. In summary, CVMP offers an appealing approach to efficient Bayesian inference for signal processing tasks, where both accuracy and execution time are important.

6 Discussion and Related Work

CVMP is a factor graph-based approximate inference method that competes with alternative factor-based methods such as Stochastic Variational Message Passing (SVMP) [9], Adaptive Importance Sampling Message Passing (AISMP) [16], Approximate Nonlinear Gaussian Message Passing (ANGMP) [7], as well as non-factor-graph-based inference methods including Black Box Variational Inference (BBVI) [20], Automatic Differentiation Variational Inference (ADVI) [3], No-U-Turn Sampler (NUTS) [2], and Conjugate-Computation Variational Inference (CVI) [4].

Following the approach used in [4, 21, 16, 9], CVMP employs Natural Gradient Descent (NGD), introduced by Amari [22, 23], and popularized by Khan as the Bayesian Learning Rule [21] to optimize the free energy functional. In contrast to Kahn's work, following Akbayrak's SVMP approach [16, 9], CVMP applies NGD locally through message updating.

BBVI, ADVI, NUTS, and CVI are slow and best for situations where accuracy is paramount, regardless of time. In contrast, since CVMP's messages are compatible with analytically computed messages in a factor graph, adding CVMP to (a few) non-linear nodes will usually have only a minor impact on execution time. As a result, adding CVMP to a hybrid message passing portfolio in a factor graph toolbox increases options for real-time inference in large models.

CVMP differs from SVMP and AISMP by utilizing a moment-matching scheme for the forward message rather than a weighted list of samples, which is also known as an "empirical distribution". The empirical distribution as a message leads to issues in interacting with closed-form messages, such as in sum-product or expectation propagation updating procedures in other parts of the factor graph. This is because the empirical distribution does not offer a convenient functional form for integration, and only moments can be derived from it. Therefore, BP and EP messages in the graph expect to interact with analytical form (e.g., Normal, Gamma) messages or require the ability to evaluate these messages at some point in the distribution domain. Empirical distributions violate both assumptions, which makes it difficult to integrate them in a hybrid message passing setting. In contrast, CVMP messages are proper analytical distributions that interface smoothly with BP and EP messages. This is an important advantage of CVMP-based inference.

Furthermore, empirical distributions restrict the model specification to scenarios where factors connected to a delta factor rely solely on the moments of the forward message from the delta factor. Consider a generative model

$$f(x_1, \ldots, x_4) = f_a(x_3, x_4, x_2) f_b(x_3, x_2) f_c(x_1, x_2) f_d(x_1).$$

In this model, f_a and f_d are arbitrary stochastic factors, while $f_b(x_3, x_2) = \delta(x_3 - h_2(x_2))$ and $f_c(x_1, x_2) = \delta(x_2 - h_1(x_1))$ are non-linear deterministic factors. It is theoretically impossible to compute the message from f_b to f_c if the message from f_c to f_b is an empirical distribution due to the ill-defined nature of division on empirical distributions since the empirical distribution is zero at most points in the distribution domain. In contrast, in CVMP, the forward message from $\delta(x_2 - h_1(x_1))$ is a proper distribution obtained through moment-matching. Consequently, the message can be computed from f_b to f_c.

CVMP can also be viewed as an extension of ANGMP to non-Gaussian cases. CVMP applies moment matching to arbitrary non-linear factors with arbitrary incoming exponential family messages, while ANGMP expects to receive mes-

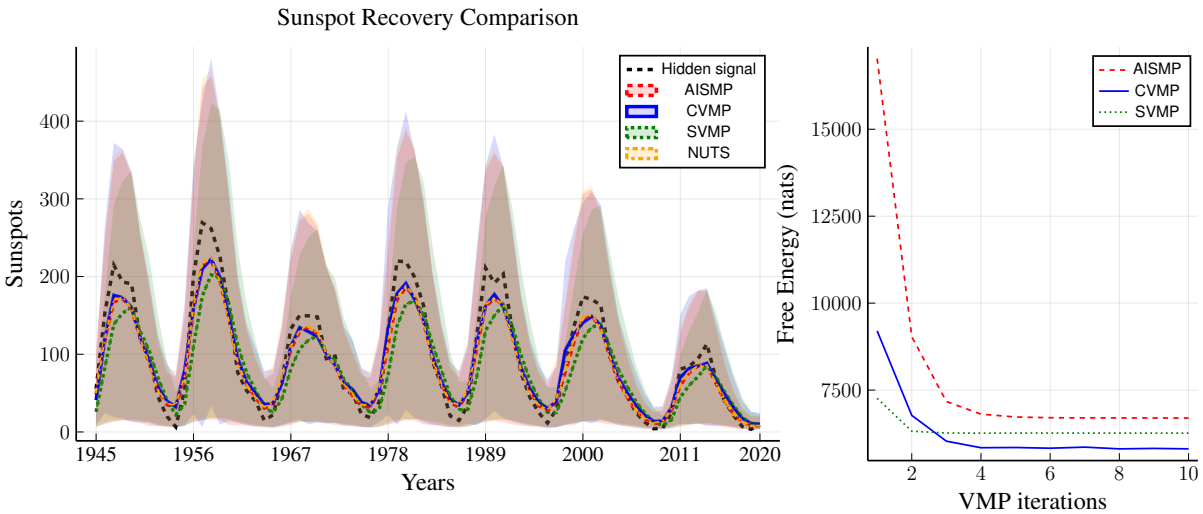

Fig. 4: Experimental validation results. Left: the black dashed line indicates hidden signal (averaged observations from [17]) The colored lines and shaded regions correspond to the mean and to the 95% confidence interval for the estimated signal of the posterior estimates $q(z_{1:T})$. Posterior estimates are color-coded based on the legend corresponding to CVMP (our proposed method), SVMP, AISMP and NUTS. Right: free energy vs inference iterations for CVMP, SVMP and AISMP algorithms. NUTS does not support Free Energy calculations.

sages in a Gaussian form. In the Gaussian case, the primary difference between CVMP and ANGMP is their moment-matching approach. ANGMP performs moment-matching using a closed-form solution, while CVMP employs a natural gradient-based method. Consequently, for quadratic non-linearities, ANGMP will perform faster.

7 Conclusions

With the objective to make progress toward real-time (variational) Bayesian inference for signal processing problems, we introduced CVMP, a novel message update rule for forward messages through non-linear nodes in a factor graph. CVMP advances state-of-the-art methods in terms of free energy minimization capabilities.

Acknowledgments

This publication is part of the ROBUST project (KICH3.LTP.20.006), which is (partly) financed by the Dutch Research Council (NWO), GN Hearing, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under program LTP KIC 2020-2023.

We are grateful for insightful discussions with and valuable implementation suggestions from Dmitry Bagaev and other BIASlab members.

8 References

[1] E.T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, 2003.

[2] Matthew D Hoffman and Andrew Gelman, "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of machine learning research*, pp. 1593–1623, 2014.

[3] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei, "Automatic Differentiation Variational Inference," *J. of Machine Learning Research*, pp. 430–474, 2017.

[4] Mohammad Khan and Wu Lin, "Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models," in *Proc. 20th Int'l Conf. on AI and Statistics*, 2017.

[5] Hans-Andrea Loeliger, Justin Dauwels, Junli Hu, Sascha Korl, Li Ping, and Frank R. Kschischang, "The Factor Graph Approach to Model-Based Signal Processing," *Proc.s of the IEEE*, 2007.

[6] İsmail Şenöz, Thijs van de Laar, Dmitry Bagaev, and Bert de Vries, "Variational Message Passing and Local Constraint Manipulation in Factor Graphs," *Entropy*, p. 807, 2021.

[7] E. Petersen, C. Hoffmann, and P. Rostalski, "On Approximate Nonlinear Gaussian Message Passing on Factor Graphs," in *2018 IEEE Statistical Signal Processing Workshop (SSP)*, 2018, pp. 513–517.

[8] David A Knowles and Thomas P Minka, "Non-conjugate Variational Message Passing for Multinomial and Binary Regression," in *Advances in NIPS*, 2011.

[9] Semih Akbayrak, İsmail Şenöz, Alp Sarı, and Bert de Vries, "Probabilistic programming with stochastic variational message passing," *Int'l J. of Approximate Reasoning*, 2022.

[10] Jonathan S. Yedidia, W.T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on Information Theory*, 2005.

[11] Thomas Minka, "Divergence Measures and Message Passing," Tech. Rep., Microsoft Research, 2005.

[12] Giovanni Di Gennaro, Amedeo Buonanno, and Francesco A. N. Palmieri, "Optimized realization of Bayesian networks in reduced normal form using latent variable model," *Soft Computing*, pp. 7029–7040, 2021.

[13] Jonathan S Yedidia, William T Freeman, and Yair Weiss, "Bethe free energy, Kikuchi approximations, and belief propagation algorithms," *Advances in NIPS*, p. 24, 2001.

[14] Martin J. Wainwright and Michael I. Jordan, "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, pp. 1–305, 2008.

[15] Ronald J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, pp. 229–256, 1992.

[16] Semih Akbayrak, Ismail Senoz, and Bert de Vries, "Adaptive Importance Sampling Message Passing," in *2022 IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, 2022.

[17] SILSO World Data Center, "The International Sunspot Number," *International Sunspot Number Monthly Bulletin and online catalogue*, 1945–2020.

[18] Dmitry Bagaev, Albert Podusenko, and Bert De Vries, "RxInfer: A Julia package for reactive real-time Bayesian inference," *Journal of Open Source Software*, p. 5161, 2023.

[19] Mike Innes, "Flux: Elegant machine learning with Julia," *Journal of Open Source Software*, p. 602, 2018.

[20] Rajesh Ranganath, Sean Gerrish, and David Blei, "Black Box Variational Inference," in *Proc. 17th Int'l Conf. on AI and Statistics*, Reykjavik, Iceland, 2014.

[21] Mohammad Emtiyaz Khan and Håvard Rue, "The Bayesian Learning Rule," 2022, arXiv:2107.04562 [cs, stat].

[22] Shun-ichi Amari, *Information Geometry and Its Applications*, Applied Mathematical Sciences. Springer Japan, 2016.

[23] Shun-ichi Amari, "Natural Gradient Works Efficiently in Learning," *Neural Computation*, pp. 251–276, 1998.