Adaptive Importance Sampling Message Passing

Semih Akbayrak, İsmail Şenöz, Bert de Vries
Eindhoven University of Technology
Eindhoven, The Netherlands
{s.akbayrak, i.senoz, bert.de.vries}@tue.nl

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834628

***Abstract*—The aim of Probabilistic Programming (PP) is to automate inference in probabilistic models. One efficient realization of PP-based inference concerns variational message passing-based (VMP) inference in a factor graph. VMP is efficient but in principle only leads to closed-form update rules in case the model consists of conjugate and/or conditionally conjugate factor pairs. Recently, Extended Variational Message Passing (EVMP) has been proposed to broaden the applicability of VMP by importance sampling-based particle methods for non-linear and non-conjugate factor pairs. EVMP automates the importance sampling procedure by employing forward messages as proposal distributions, which unfortunately may lead to inaccurate estimation results and numerical instabilities in case the forward message is not a good representative of the unknown correct posterior. This paper addresses this issue by integrating an *adaptive* importance sampling procedure with message passing-based inference. The resulting method is a hyperparameter-free approximate inference engine that combines recent advances in stochastic adaptive importance sampling and optimization methods. We provide an implementation for the proposed method in the Julia package *ForneyLab.jl*.**

***Index Terms*—approximate Bayesian inference, importance sampling, message passing, variational inference**

I. Introduction

Inference is often considered the challenging stage of probabilistic modelling as it requires expertise in (approximate) Bayesian inference methods. Probabilistic Programming Languages (PPLs) [1] aim to automate the inference stage so that end-users can focus only on model development [2]–[4]. However, achieving this goal is also challenging as it necessitates automatable and broadly applicable inference algorithms that are hopefully hyperparameter-free, too.

This paper proposes a broadly applicable, hyperparameter-free inference algorithm called Adaptive Importance Sampling Message Passing (AIS-MP). AIS-MP is a hybrid Monte Carlo message passing-based inference approach that combines the efficiency and the speed of rule-based message passing algorithms , such as Belief Propagation (BP) [5], [6], Variational Message Passing (VMP) [7], [8], and Expectation Propagation (EP) [9], [10] with the generality of Monte Carlo sampling on Forney-style Factor Graphs (FFGs).

Our work closely relates to the *Extended* Variational Message Passing (EVMP) algorithm [11], which extends the applicability of VMP to non-conjugate and non-linear models. EVMP achieves this through estimation of analytically intractable expectation quantities in VMP message calculations, either through a Laplace approximation [12, Section 4.4] or through importance sampling (IS) [13], [14]. To reduce the burden on PPL end users to specify hyperparameter values and proposal distributions, EVMP casts so-called *forward messages* as proposal distributions in IS. This method coincides with the popular Bootstrap particle filtering approach [15], [16], but unfortunately, the method suffers from imprecise expectation estimations and numerical instabilities if the forward message is not a good representative of the correct posterior distribution.

AIS-MP approaches the above shortcomings of EVMP with an *adaptive* IS [17] procedure. Specifically, AIS-MP initializes the proposal distribution with a forward message and runs a stochastic optimization to tune this distribution iteratively until the number of efficient samples exceeds a certain threshold. In the stochastic optimization procedure of the proposal distribution, we use an approach introduced in Stochastic Gradient Population Monte Carlo (SG-PMC) [18], by generalizing it to the exponential family of distributions, similar to [19], with an α-divergence [20] cost function, where $\alpha = 2$. We provide an implementation of AIS-MP in a Julia [21] language-based PPL, *ForneyLab.jl* [22] and demonstrate its performance on a non-conjugate Gamma state-space model.

II. Background

In this section, we briefly summarize Forney-style Factor Graphs (FFGs) [23] and Variational Message Passing (VMP) on FFGs. An FFG is a probabilistic graphical model comprised of factor nodes and edges that are associated with conditional distributions and random variables, respectively. Random variables that are argument to more than two factors branch out through equality nodes in FFGs (see Figure 2).

Assume a probabilistic model $f(\boldsymbol{y}, \boldsymbol{z})$ for a given set of observations $\boldsymbol{y} = \{y_1, y_2, ..., y_N\}$ and hidden variables $\boldsymbol{z} = \{z_1, z_2, ..., z_M\}$. In case exact inference is intractable, the variational inference method approximates the exact posterior $p(\boldsymbol{z}|\boldsymbol{y})$ by a "recognition" distribution $q(\boldsymbol{z})$ through minimization of the (variational) Free Energy

$$\mathcal{F}[q(\boldsymbol{z})] = \mathbb{E}_{q(\boldsymbol{z})}\left[\log q(\boldsymbol{z}) - \log f(\boldsymbol{y}, \boldsymbol{z})\right] , \tag{1}$$

where $\mathbb{E}_{q(\boldsymbol{z})}[\cdot]$ refers to expectation with respect to $q(\boldsymbol{z})$. To cast the free energy minimization as an iterative coordinate-descent optimization procedure, $q(\boldsymbol{z})$ is often chosen among factorized distribution families [12].

Consider the sub-graph given in Figure 1a with a recognition distribution consisting of factors $q(z_k)q(\boldsymbol{z_{a\setminus k}})q(\boldsymbol{z_{b\setminus k}})$. Coordinate-descent optimization of the free energy in this factorized graph is achieved through a distributed inference

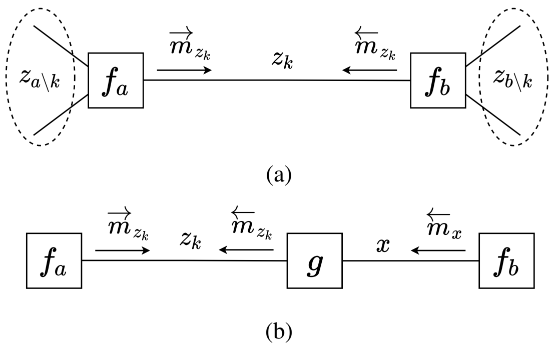

Fig. 1: (a) A sub-graph with factor nodes f_a and f_b connected through z_k. (b) A deterministic node $\delta(x - g(z_k))$ allows us to specify complex models.

procedure, called Variational Message Passing (VMP) [7]. In an FFG setting, the VMP update for latent variable z_k is described by [8]

$$\overrightarrow{m}_{z_k}(z_k) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{z_{a\backslash k}})}[\log f_a(\boldsymbol{z_a})]\right) \tag{2a}$$

$$\overleftarrow{m}_{z_k}(z_k) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{z_{b\backslash k}})}[\log f_b(\boldsymbol{z_b})]\right) \tag{2b}$$

$$q(z_k) = \overrightarrow{m}_{z_k}(z_k)\overleftarrow{m}_{z_k}(z_k) \Big/ \int \overrightarrow{m}_{z_k}(z_k)\overleftarrow{m}_{z_k}(z_k)\mathrm{d}z_k\,, \tag{2c}$$

where $\boldsymbol{z_a}$ denotes the arguments of the factor f_a, $\boldsymbol{z_{a\backslash k}}$ stands for all the arguments of f_a but z_k, and $\overrightarrow{m}_{z_k}(z_k)$ and $\overleftarrow{m}_{z_k}(z_k)$ are respectively forward and backward messages.

In practice, one has to specify the probabilistic model carefully such that the messages in (2) and the marginal posterior in (2c) can easily be calculated. A natural way of satisfying these conditions is to choose factors as conjugate (or conditionally conjugate) pairs that leads to following messages

$$\overrightarrow{m}(z_k) \propto \exp\left(\overrightarrow{\phi}_{z_k}(z_k)^\intercal \cdot \overrightarrow{\eta}_{z_k}\right) \tag{3a}$$

$$\overleftarrow{m}(z_k) \propto \exp\left(\overleftarrow{\phi}_{z_k}(z_k)^\intercal \cdot \overleftarrow{\eta}_{z_k}\right), \tag{3b}$$

where $\overrightarrow{\phi}_{z_k}(z_k) = \overleftarrow{\phi}_{z_k}(z_k) = \phi_{z_k}(z_k)$ since f_a and f_b are conjugate factor pairs. Substituting (3) in (2c), the approximate posterior turns out to be

$$q(z_k) = h_{z_k}(z_k)\exp\left(\phi_{z_k}(z_k)^\intercal \cdot \underbrace{(\overrightarrow{\eta}_{z_k} + \overleftarrow{\eta}_{z_k})}_{\eta_{z_k}} - A_{z_k}(\eta_{z_k})\right),$$

which is a member of exponential family of distributions [24] with constant base measure $h_{z_k}(z_k)$, sufficient statistics $\phi_{z_k}(z_k)$, natural parameters η_{z_k} and log-partition function $A_{z_k}(\eta_{z_k})$. If the underlying graph consists of conditionally conjugate factors then VMP is a very efficient algorithm for approximate Bayesian inference. The presence of non-conjugate factor pairs often prevents efficient realization of VMP in practice.

Extended Variational Message Passing (EVMP) [11] removes the limitations of VMP by estimating the expectation quantities that appear in VMP messages by importance sampling (IS) in an automated way. Consider Figure 1b. This time we insert a deterministic mapping $\delta(x - g(z_k))$ between the factors f_a and f_b, which enables the end-user to specify more complex models using deterministic functions $g(z_k)$. In this sub-graph, the message from the deterministic node to z_k is

$$\begin{aligned}\overleftarrow{m}_{z_k}(z_k) &= \int \overleftarrow{m}_x(x)\delta(x - g(z_k))\mathrm{d}x \\ &= \overleftarrow{m}_x(g(z_k)) \propto \exp\left(\overleftarrow{\phi}_x(g(z_k))^\intercal \cdot \overleftarrow{\eta}_x\right),\end{aligned} \tag{4}$$

which often leads to a backward message $\overleftarrow{m}(z_k)$ that differs from the forward message $\overrightarrow{m}(z_k)$ in its sufficient statistics. In this case, we are often prevented from calculating the approximate marginal $q(z_k)$ analytically, since the normalization factor in (2c) is not available in closed form. As a remedy, EVMP introduces an additional approximation in the calculation of the posterior $p(z_k|\boldsymbol{y})$, leading to

$$q(z_k) \approx \tilde{q}(z_k) = \sum_{i=1}^{N} w_{z_k}^{(i)}\delta(z_k - z_k^{(i)}), \tag{5}$$

$$\text{where } z_k^{(i)} \sim \overrightarrow{m}_{z_k}(z_k), w_{z_k}^{(i)} = \frac{\overleftarrow{m}_{z_k}(z_k^{(i)})}{\sum_{i=1}^{N}\overleftarrow{m}_{z_k}(z_k^{(i)})}.$$

Similarly, $q(x)$ is represented by

$$q(x) \approx \tilde{q}(x) = \sum_{i=1}^{N} w_{z_k}^{(i)}\delta(x - g(z_k^{(i)})).$$

The above approximations follow from IS with a proposal distribution $\overrightarrow{m}_{z_k}(z_k)$. Once $q(z_k)$ and $q(x)$ are represented with weighted samples, EVMP estimates the expectations, such as $\mathbb{E}_{q(z_k)}[\Phi(z_k)]$ and $\mathbb{E}_{q(x)}[\Phi(x)]$ for an arbitrary function $\Phi(\cdot)$, that are required in calculation of VMP messages around f_a and f_b with Monte Carlo summations, e.g.,

$$\mathbb{E}_{q(x)}[\Phi(x)] \approx \sum_{i=1}^{N} w_{z_k}^{(i)}\Phi(g(z_k^{(i)})),$$

given that the support of $\overrightarrow{m}_{z_k}(z_k)$ encapsulates the support of $q(z_k)$ [16, Page 118]. Casting $\overrightarrow{m}_{z_k}(z_k)$ as the proposal distribution for IS obviates the need for proposal distribution specification and hence allows EVMP to be automated in message passing-based PPLs. However, this automated process sometimes entails imprecise estimations when the proposal distribution is not a good representative of the unknown posterior. Next, we will improve the performance of EVMP by adaptively adjusting proposal distributions in IS.

III. AIS-MP

In the previous section, we showed that EVMP employs the pre-defined functional forms of the VMP messages for inference and fills in the expectation quantities required in message calculations with their estimates calculated via IS. In this section, we present Adaptive Importance Sampling Message Passing (AIS-MP) that aims to improve the IS procedure of EVMP by using better proposal distributions.

A. Adaptive IS with Stochastic Gradient Descent

Consider Figure 1 again. We define a weighted particle approximation $\tilde{q}(z_k)$ as

$$q(z_k) \approx \tilde{q}(z_k) = \sum_{i=1}^{N} w_{z_k}^{(i)} \delta(z_k - z_k^{(i)}) \,, \tag{6}$$

where

$$z_k^{(i)} \sim \pi(z_k),\, w_{z_k}^{(i)} = \frac{\frac{\overrightarrow{m}_{z_k}(z_k^{(i)})\overleftarrow{m}_{z_k}(z_k^{(i)})}{\pi(z_k^{(i)})}}{\sum_{j=1}^{N} \frac{\overrightarrow{m}_{z_k}(z_k^{(j)})\overleftarrow{m}_{z_k}(z_k^{(j)})}{\pi(z_k^{(j)})}}.$$

This time the proposal distribution $\pi(z_k)$ explicitly appears in the computation of weights (6), since we do not set $\pi(z_k) = \overrightarrow{m}_{z_k}(z_k)$. In selection of optimal $\pi(z_k)$, we choose to find a minimum variance, unbiased estimator of the normalization constant of $q(z_k)$ that is $\int \overrightarrow{m}_{z_k}(z_k)\overleftarrow{m}_{z_k}(z_k)\mathrm{d}z_k$. As shown in [20], this can be achieved by minimizing the α-divergence between $\overrightarrow{m}_{z_k}(z_k)\overleftarrow{m}_{z_k}(z_k)$ and $\pi(z_k)$ for $\alpha = 2$:

$$\begin{aligned} D_2[q(z_k)||\pi(z_k)] &= \frac{1}{2}\int \frac{\left(\overrightarrow{m}_{z_k}(z_k)\overleftarrow{m}_{z_k}(z_k) - \pi(z_k)\right)^2}{\pi(z_k)}\mathrm{d}z_k \\ &\propto \int \frac{q(z_k)^2}{\pi(z_k)}\mathrm{d}z_k = \mathbb{E}_{q(z_k)}\left[\frac{q(z_k)}{\pi(z_k)}\right], \end{aligned} \tag{7}$$

where the multiplicative and additive constants are dropped. The last line follows from that we choose our proposal $\pi(z_k)$ to be a proper distribution. More precisely, we constrain $\pi(z_k)$ to be in the same distribution family with $\overrightarrow{m}_{z_k}(z_k)$, i.e.,

$$\pi(z_k;\lambda) = \overrightarrow{h}_{z_k}(z_k)\exp\left(\overrightarrow{\phi}_{z_k}(z_k)^{\intercal}\lambda - \overrightarrow{A}_{z_k}(\lambda)\right), \tag{8}$$

with a constant $\overrightarrow{h}_{z_k}(z_k)$. Having specified the functional form of $\pi(z_k;\lambda)$ in an exponential family, we shall iteratively tune its parameters in such a way that $D_2[q(z_k)||\pi(z_k;\lambda)]$ is minimized:

$$\lambda^{(t)} \longleftarrow \lambda^{(t-1)} - \rho^{(t)}\nabla_\lambda D_2[q(z_k)||\pi(z_k;\lambda)], \tag{9}$$

where t denotes the iteration index and $\rho^{(t)}$ is the step size at iteration t. We obtain $\nabla_\lambda D_2[q(z_k)||\pi(z_k;\lambda)]$ by

$$\begin{aligned} \nabla_\lambda D_2 &= -\mathbb{E}_{q(z_k)}\left[\frac{q(z_k)\nabla_\lambda \pi(z_k)}{\pi(z_k)^2}\right] \\ &= -\mathbb{E}_{q(z_k)}\left[\frac{q(z_k)}{\pi(z_k)}\nabla_\lambda \log \pi(z_k)\right] \\ &= -\mathbb{E}_{q(z_k)}\left[\frac{q(z_k)}{\pi(z_k)}(\overrightarrow{\phi}_{z_k}(z_k) - \mathbb{E}_\pi[\overrightarrow{\phi}_{z_k}(z_k)])\right]. \end{aligned} \tag{10}$$

The second line follows from $\frac{\nabla_\lambda \pi(z_k)}{\pi(z_k)} = \nabla_\lambda \log \pi(z_k)$ [2]. The last line is due to the property of exponential family of distributions that the gradient of the log-normalizer is expectation of sufficient statistics [24], i.e., $\nabla_\lambda \overrightarrow{A}_{z_k}(\lambda) = \mathbb{E}_\pi[\overrightarrow{\phi}_{z_k}(z_k)]$, which is available in closed-form. However, the overall expectation required to calculate $\nabla_\lambda D_2[q||\pi]$ does not have an analytical solution since $q(z_k)$ is unknown. Instead, we follow SG-PMC's stochastic approximation approach [18] to estimate the true gradient with

$$\begin{aligned} \tilde{\nabla}_\lambda D_2 &= -\mathbb{E}_{\tilde{q}(z_k)}\left[\frac{q(z_k)}{\pi(z_k)}(\overrightarrow{\phi}_{z_k}(z_k) - \mathbb{E}_\pi[\overrightarrow{\phi}_{z_k}(z_k)])\right] \\ &= -\sum_{i=1}^{N} w_{z_k}^{(i)} \frac{q(z_k^{(i)})}{\pi(z_k^{(i)})}(\overrightarrow{\phi}_{z_k}(z_k^{(i)}) - \mathbb{E}_\pi[\overrightarrow{\phi}_{z_k}(z_k)]). \end{aligned} \tag{11}$$

Notice that $q(z_k^{(i)}) \propto \overrightarrow{m}_{z_k}(z_k^{(i)})\overleftarrow{m}_{z_k}(z_k^{(i)})$, hence using the weighting definition in (6) we can write

$$\frac{q(z_k^{(i)})}{\pi(z_k^{(i)})} \propto w_{z_k}^{(i)}. \tag{12}$$

We now substitute (12) back in (11) and find a noisy gradient estimate of $\nabla_\lambda D_2[q||\pi]$ in closed-form:

$$\tilde{\nabla}_\lambda D_2 \propto -\sum_{i=1}^{N} {w_{z_k}^{(i)}}^2 \left(\overrightarrow{\phi}_{z_k}(z_k^{(i)}) - \mathbb{E}_\pi[\overrightarrow{\phi}_{z_k}(z_k)]\right). \tag{13}$$

Substituting $\nabla_\lambda D_2[q||\pi]$ with a noisy gradient estimate $\tilde{\nabla}_\lambda D_2[q||\pi]$ in (9), and setting $\rho^{(t)}$ according to Robins-Monro conditions [25], i.e., $\sum_{t=1}^{\infty} \rho^{(t)} = \infty$, $\sum_{t=1}^{\infty} {\rho^{(t)}}^2 < \infty$, we get a stochastic gradient descent procedure to tune the parameters λ of the proposal distribution $\pi(z_k;\lambda)$.

In our optimization strategy, we use $\overrightarrow{m}_{z_k}(z_k)$ as the initial proposal distribution $\pi(z_k;\lambda^{(0)})$, i.e., $\lambda^{(0)} = \overrightarrow{\eta}_{z_k}$ and iteratively refine it. At the end of iteration t, we collect new weighted particles to be used in gradient estimation (13) at iteration $t+1$ by employing $\pi(z_k;\lambda^{(t)})$ in (6).

To diagnose the convergence of the stochastic approximation, we keep track of the number of efficient particles [16, Chapter 7]:

$$n_{\text{eff}} = 1 \Big/ \sum_{i=1}^{N} {w_{z_k}^{(i)}}^2. \tag{14}$$

Once the number of efficient particles exceeds the specified threshold, e.g., $n_{\text{eff}} > N/10$ [16, Page 124], we stop the stochastic approximation procedure and use the converged $\pi(z_k)$ in (6) to evaluate $\tilde{q}(z_k)$. This procedure relieves the end-user from choosing the number of iterations and carries out the convergence diagnosis automatically.

B. Backward Message Calculation with Moment Matching

Approximating $q(z_k)$ by a set of weighted samples $\tilde{q}(z_k)$ suffices to execute EVMP. We can also find an approximation $\bar{q}(z_k)$ within the distribution family of $\overrightarrow{m}_{z_k}(z_k)$ by using the weighted samples $\tilde{q}(z_k)$ and moment matching [9]:

$$\bar{q}(z_k) \propto \exp\left(\overrightarrow{\phi}_{z_k}(z_k)^{\intercal} \underbrace{\psi^{-1}\left(\left[\mathbb{E}_{\tilde{q}(z_k)}[z_k], \mathbb{V}_{\tilde{q}(z_k)}[z_k]\right]^{\intercal}\right)}_{\bar{\eta}_{z_k}}\right).$$

Here, $\mathbb{V}_{\tilde{q}(z_k)}[z_k]$ is the variance of z_k calculated over $\tilde{q}(z_k)$ and $\psi(\cdot)$ is a mapping from natural parameters to central moments for the chosen exponential family distribution $\bar{q}(z_k)$, i.e.,

$$\psi(\eta_{z_k}) = \left[\mathbb{E}_{\bar{q}(z_k)}[z_k], \mathbb{V}_{\bar{q}(z_k)}[z_k]\right]^{\intercal}. \tag{15}$$

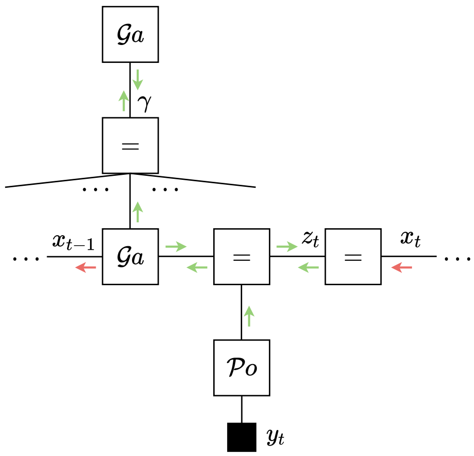

Fig. 2: A time slice of the FFG we build in ForneyLab for Gamma state-space model. The equality node that generates an auxilary variable x_t performs AIS-MP and approximates a non-Gamma message (shown by red arrow) with a Gamma message (shown by green).

The advantages of moment matching are twofold. Firstly, in the free energy calculation, $\bar{q}(z_k)$ yields a closed form solution for the entropy term, corresponding to z_k. Secondly, moment matching allows us to approximate the backward message $\overleftarrow{m}_{z_k}(z_k)$ with $\overleftarrow{\nu}_{z_k}(z_k)$ by dividing $\bar{q}(z_k)$ with $\overrightarrow{m}_{z_k}(z_k)$ [9], [26]:

$$\overleftarrow{\nu}_{z_k}(z_k) \propto \exp\left(\overrightarrow{\phi}_{z_k}(z_k)^\intercal(\bar{\eta}_{z_k} - \overrightarrow{\eta}_{z_k})\right). \tag{16}$$

Apart from being employed in VMP seamlessly, the above message is likely to be in a convenient functional form to be integrated with BP or EP.

C. Algorithm and Node-level Implementation

AIS-MP is summarized in Algorithm 1 and implemented in a Julia language-based message passing PPL *ForneyLab.jl* (Available at https://github.com/semihakbayrak/ForneyLab.jl/tree/AIS-MP). By default, the number of samples is set to 1000 and ADAM optimizer [27] from Flux.jl [28] package is employed to adaptively adjust step sizes. The end-user is free to change these hyperparameters. Note that we keep track of the convergence and automatically determine whether to terminate the optimization.

If a deterministic relation is not needed in the model specification but the inference is still challenging, the end-user of *ForneyLab.jl* can execute AIS-MP by introducing an auxilary random variable $x = z_k$.

IV. Related Work

AIS-MP is an instance of the class of approximate inference methods for probabilistic programming, like Black Box Variational Inference (BBVI) [2], Automatic Differentiation Variational Inference (ADVI) [3] and No-U-Turn Sampler (NUTS) [4]. Unlike BBVI, ADVI and NUTS, AIS-MP utilizes stochastic approximation methods only when closed form message passing algorithms do not suffice to run inference in non-conjugate and nonlinear sections of the model specification. Similar hybrid approaches are proposed in [30], [31]. We differ from them in that AIS-MP estimates expectation quantities with IS, which is accompanied by the number of efficient samples to track the convergence of stochastic approximations.

Adaptive importance sampling has been incorporated to enhance the performance of variational inference in [32]. Our work differs from theirs in several notable ways. Most notably, they utilize adaptive importance sampling to reduce the variance of the free energy gradient estimates in BBVI. Whereas we use adaptive importance sampling directly in the approximation of the posterior marginals and the messages. Secondly, they use Monte Carlo moment matching in approximation of the optimal proposal distribution for free energy gradient estimation. In contrast, we adhere to SG-PMC's stochastic optimization approach to tune proposal distributions by generalizing it to exponential family of distributions and minimizing α-divergence for $\alpha = 2$.

The gradient estimate in (13) substantially coincides with the noisy gradient derived in [19] except that their procedure is in a fully online setting (no summation term as in (13)). Another difference is that they minimize the variance of the estimator for expectation quantities such as $\mathbb{E}_{q(z_k)}[\Phi(z_k)]$, whereas we minimize the variance of the estimator for the normalization constant of $q(z_k)$ by aiming to get a good weighted samples representation $\tilde{q}(z_k)$.

Algorithm 1 AIS-MP around a deterministic node in an FFG

Require: A deterministic node $\delta(x - g(z_k))$
Collect $\overrightarrow{m}_{z_k}(z_k)$, $\overleftarrow{m}_{z_k}(z_k) = \overleftarrow{m}_x(g(z_k))$
Set $t = 0$; $\pi(z_k; \lambda^{(0)}) = \overrightarrow{m}_{z_k}(z_k)$
Find $\tilde{q}(z_k)$ using (6)
while $n_{\text{eff}} < N/10$ **do** ▷ n_{eff} as (14), $N = 1000$ by default
  $t += 1$; set $\rho^{(t)}$ ▷ ADAM optimizer by default
  Run (9) using (13)
  Find $\tilde{q}(z_k)$ using (6)
Find $\bar{q}(z_k)$ and $\overleftarrow{\nu}_{z_k}(z_k)$ using (15) and (16)
Set $\tilde{q}(x) = \sum_{i=1}^{N} w_{z_k}^{(i)} \delta(x - g(z_k^{(i)}))$

V. Experimental Validation

In this section, we use AIS-MP in *ForneyLab.jl* to analyze the yearly solar activities between 1945 and 2020 over a sunspots data set (Source: WDC-SILSO, Royal Observatory of Belgium, Brussels [29]; see Figure 3; experiments available at https://github.com/biaslab/AIS-MP). Data samples are rational numbers as they are calculated by averaging count data. To reflect the count nature of the data, we round data sample values to their closest integer values and model them with Poisson likelihoods. We designed a non-conjugate Gamma state-space model to track the rate parameters of the Poisson likelihoods.

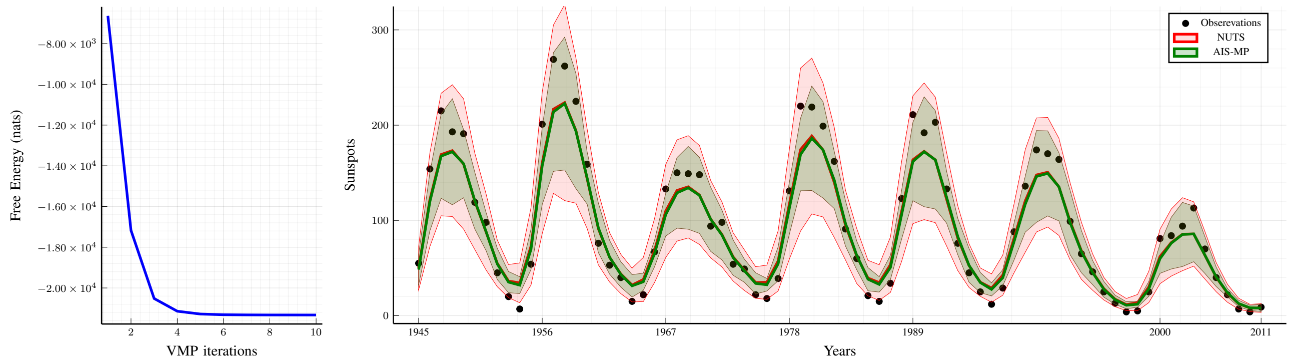

Fig. 3: Figure summarizes the results of the experimental validation. On the left, free energy over VMP iterations are visualized for AIS-MP algorithm. On the right, black dots indicate sunspot observations [29] rounded to closest integer values. The lines and shaded regions correspond to mean and variance of the posterior estimates $q(z_t)$. Posterior estimates are color-coded based on the legend corresponding to AIS-MP (this paper) and NUTS (baseline) [4].

More precisely, we propose the following generative model for the sunspots dataset:

$$p(\boldsymbol{y},\boldsymbol{z},\gamma) = p(\gamma)p(z_1|\gamma)p(y_1|z_1)\prod_{t=2}^{T} p(z_t|z_{t-1},\gamma)p(y_t|z_t)$$

$$\begin{aligned} \text{where} \qquad p(\gamma) &= \mathcal{G}a(\gamma; 1000, 1) \\ p(z_1|\gamma) &= \mathcal{G}a(z_1; 1, \gamma) \\ p(z_t|z_{t-1},\gamma) &= \mathcal{G}a(z_t; z_{t-1}, \gamma) \\ p(y_t|z_t) &= \mathcal{P}o(y_t; z_t), \end{aligned}$$

where $\mathcal{G}a(\cdot; a, b)$ denotes Gamma distribution with shape a and rate b, and $\mathcal{P}o(\cdot;\zeta)$ is Poisson distribution with rate ζ. We run VMP on the model by utilizing IS to estimate expectations quantities that are not available in closed form. We assumed a mean-field factorization on the recognition distribution

$$q(\gamma,\boldsymbol{z}) = q(\gamma)\prod_{t=1}^{T} q(z_t). \tag{17}$$

This is a challenging model specification for EVMP as the chosen priors lead to forward VMP messages that significantly diverge from the unknown correct posteriors. Hence, we run AIS-MP to automatically tune the proposal distributions by IS estimates of expectations. We build an FFG as in Figure 2 in *ForneyLab.jl*. Note that we introduce deterministic equality nodes that generate dummy variables $\boldsymbol{x} = \boldsymbol{z}$ and perform AIS-MP around these nodes. Running VMP for 10 iterations, the free energy converges as in Figure 3 (left) and we get Gamma approximate distributions $q(z_t)$, mean and variance of which are visualized in Figure 3 (right).

We compare AIS-MP's estimates with NUTS's in Figure 3. We use Turing [33] probabilistic programming package of Julia language to run the NUTS inference engine. We observe that the mean estimates substantially coincide, whereas NUTS's variance estimates are larger in comparison to AIS-MP's. The difference in the variance estimations is not surprising as we use a fully factorized distribution to perform approximate inference in the AIS-MP case, whereas NUTS performs inference over the joint distribution of the random variables. In terms of run time, NUTS is preferable to AIS-MP for this model. AIS-MP converges in 6 VMP iterations, which takes roughly 2.5 minutes to execute in *ForneyLab.jl* including graph construction, whereas NUTS converges very fast with a reverse mode automatic differentiation [34], in less than 3 seconds in our personal computer. Nevertheless, AIS-MP can still be a good alternative to NUTS in different model specifications. For example, Switching State-Space Model (SSSM) variants [35] comprise both continuous and discrete variables, hence NUTS must be combined with other samplers that perform inference for discrete variables, which sometimes does not yield satisfactory estimations (see [11, Section 4.3]). As opposed to NUTS, AIS-MP can be used to estimate discrete variables. For an SSSM example, we provide an AIS-MP implementation in our experiments repository. In the SSSM example, forward messages yield good proposal distributions and AIS-MP executes EVMP in effect without the need for stochastic optimization. We additionally provide a simple Categorical-Normal experiment to demonstrate how AIS-MP differs from EVMP by running stochastic optimization to estimate discrete variables.

VI. Discussion and Conclusion

In this paper, we propose Adaptive Importance Sampling Message Passing (AIS-MP) that uses a stochastic adaptive importance sampling approach to estimate the required expectations in the approximation of messages in FFGs. AIS-MP aims to mitigate the shortcomings of the previously proposed Extended VMP (EVMP) algorithm for automated VMP in message passing-based PPLs. As opposed to EVMP, AIS-MP consists of a stochastic optimization procedure, and hence inference is slower compared to EVMP. Nonetheless, as demonstrated by experimental validation, AIS-MP performs better inference on models that EVMP cannot handle. We coded AIS-MP in the Julia language-based PPL, *ForneyLab.jl* and aim to release it as a full inference engine in the future.

References

[1] Jan-Willem van de Meent, Brooks Paige, Hongseok Yang, and Frank Wood. An Introduction to Probabilistic Programming. *arXiv:1809.10756 [cs, stat]*, September 2018. arXiv: 1809.10756.

[2] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

[3] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017. Publisher: JMLR. org.

[4] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[5] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[6] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[7] John Winn and Christopher M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(Apr):661–694, 2005.

[8] Justin Dauwels. On Variational Message Passing on Factor Graphs. In *IEEE International Symposium on Information Theory*, pages 2546–2550, June 2007.

[9] Thomas P Minka. Expectation Propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

[10] Aki Vehtari, Andrew Gelman, Tuomas Sivula, Pasi Jylänki, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P Cunningham, David Schiminovich, and Christian P Robert. Expectation Propagation as a Way of Life: A Framework for Bayesian Inference on Partitioned Data. *Journal of Machine Learning Research*, 21(17):1–53, 2020.

[11] Semih Akbayrak, Ivan Bocharov, and Bert de Vries. Extended Variational Message Passing for Automated Approximate Bayesian Inference. *Entropy*, 23(7):815, June 2021.

[12] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[13] Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010. Publisher: Wiley Online Library.

[14] Víctor Elvira and Luca Martino. Advances in Importance Sampling. *arXiv:2102.05407 [stat]*, March 2022. arXiv: 2102.05407.

[15] Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEEE proceedings on radar and signal processing*, volume 140, pages 107–113, 1993. Issue: 2.

[16] Simo Särkkä. *Bayesian filtering and smoothing*, volume 3. Cambridge University Press, 2013.

[17] Monica F. Bugallo, Victor Elvira, Luca Martino, David Luengo, Joaquin Miguez, and Petar M. Djuric. Adaptive Importance Sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, July 2017. Conference Name: IEEE Signal Processing Magazine.

[18] Yousef El-Laham and Mónica F. Bugallo. Stochastic Gradient Population Monte Carlo. *IEEE Signal Processing Letters*, 27:46–50, 2020. Conference Name: IEEE Signal Processing Letters.

[19] Ernest K. Ryu and Stephen P. Boyd. Adaptive Importance Sampling via Stochastic Convex Programming. *arXiv:1412.4845 [math, stat]*, January 2015. arXiv: 1412.4845.

[20] Thomas Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Cambridge, 2005.

[21] Jeff Bezanson, Stefan Karpinski, Viral B Shah, and Alan Edelman. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

[22] Marco Cox, Thijs van de Laar, and Bert de Vries. A factor graph approach to automated design of Bayesian signal processing algorithms. *International Journal of Approximate Reasoning*, 104:185–204, January 2019.

[23] Hans-Andrea Loeliger, Justin Dauwels, Junli Hu, Sascha Korl, Li Ping, and Frank R Kschischang. The factor graph approach to model-based signal processing. *Proceedings of the IEEE*, 95(6):1295–1322, 2007. Publisher: IEEE.

[24] Martin J Wainwright and Michael I Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[25] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. Publisher: Institute of Mathematical Statistics.

[26] Marco Cox and Bert De Vries. Robust Expectation Propagation in Factor Graphs Involving Both Continuous and Binary Variables. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2583–2587, Rome, September 2018. IEEE.

[27] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[28] Michael Innes, Elliot Saba, Keno Fischer, Dhairya Gandhi, Marco Concetto Rudilosso, Neethu Mariya Joy, Tejan Karmali, Avik Pal, and Viral Shah. Fashionable Modelling with Flux. *arXiv:1811.01457 [cs]*, November 2018. arXiv: 1811.01457.

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

[30] Mohammad Khan and Wu Lin. Conjugate-Computation Variational Inference: Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models. In *Artificial Intelligence and Statistics*, pages 878–887. PMLR, 2017.

[31] Semih Akbayrak and Bert de Vries. Reparameterization Gradient Message Passing. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, A Coruna, Spain, September 2019. IEEE.

[32] Ximing Li, Changchun Li, Jinjin Chi, and Jihong Ouyang. Variance Reduction in Black-box Variational Inference by Adaptive Importance Sampling. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 2404–2410, Stockholm, Sweden, July 2018. International Joint Conferences on Artificial Intelligence Organization.

[33] Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: A Language for Flexible Probabilistic Inference. In *International Conference on Artificial Intelligence and Statistics*, pages 1682–1690. PMLR, March 2018.

[34] Jarrett Revels. *ReverseDiff.jl*. 2017.

[35] Zoubin Ghahramani and Geoffrey E Hinton. Variational learning for switching state-space models. *Neural computation*, 12(4):831–864, 2000. Publisher: MIT Press.