Reparameterization Gradient Message Passing

Semih Akbayrak
Eindhoven University of Technology
Eindhoven, the Netherlands
s.akbayrak@tue.nl

Bert de Vries
GN Hearing and TU Eindhoven
Eindhoven, the Netherlands
bdevries@ieee.org

***Abstract*—In this paper we consider efficient message passing based inference in a factor graph representation of a probabilistic model. Current message passing methods, such as belief propagation, variational message passing or expectation propagation, rely on analytically pre-computed message update rules. In practical models, it is often not feasible to analytically derive all update rules for all factors in the graph and as a result, efficient message passing-based inference cannot proceed. In related research on (non-message passing-based) inference, a “reparameterization trick” has lead to a considerable extension of the class of models for which automated inference is possible. In this paper, we introduce Reparameterization Gradient Message Passing (RGMP), which is a new message passing method based on the reparameterization gradient. In most models, the large majority of messages can be analytically derived and we resort to RGMP only when necessary. We will argue that this kind of hybrid message passing leads naturally to low-variance gradients.**

I. Introduction

In this paper, we focus on automating inference for probabilistic models. In particular, we extend the range of message passing-based inference methods for factor graph representations of probabilistic models [1]. In order to allow computationally efficient inference by message passing, factor graphs take advantage of factorization and independence relations in probabilistic models. For linear Gaussian models, analytically computable message update rules exist (called: belief propagation (BP)) that lead to exact Bayesian inference if the graph is a tree. For an even larger class of models, approximate Bayesian inference may be implemented by variational message passing (VMP) [2]. VMP is based on approximating the posterior for latent variable $\boldsymbol{Z}$ by an instrumental (variational) distribution $q(\boldsymbol{Z})$ and minimizing the Free Energy functional:

$$\begin{aligned} F[q] &= \mathbb{E}_{q(\boldsymbol{Z})}\left[\log\frac{q(\boldsymbol{Z})}{p(\boldsymbol{Y},\boldsymbol{Z})}\right] \\ &= -\log p(\boldsymbol{Y}) + KL(q(\boldsymbol{Z})\,||\,p(\boldsymbol{Z}|\boldsymbol{Y}))\,, \end{aligned} \tag{1}$$

where $p(\boldsymbol{Y},\boldsymbol{Z})$ is the generative model over latent variables $\boldsymbol{Z}$ and observed variables $\boldsymbol{Y}$, $q(\boldsymbol{Z})$ is a variational distribution and $KL(.||.)$ is the Kullback-Leibler divergence [3]. Alternatively, Expectation Propagation (EP) [4] is a popular third message passing technique that relies on the inclusive KL divergence $KL(p(\boldsymbol{Z}|\boldsymbol{Y})\,||\,q(\boldsymbol{Z}))$. For all of the above message passing methods, it is necessary to derive analytically computable message update rules for all factors in the model. For many models, this is not possible for all factors and in that case, efficient inference by message passing cannot proceed.

In other (non-message passing-based) research developments on automating variational inference, a considerable amount of progress has been achieved with the so-called *reparameterization trick*. If the latent random variables $\boldsymbol{Z}$ are considered outputs of differentiable, injective functions, then a “reparameterization trick” facilitates computation of noisy free energy gradients, which can be used to minimize the free energy by a stochastic optimization method, [5]–[7].

In the current paper, we formulate reparameterization gradient variational inference as message passing in (Forney-style) factor graphs [1]. In a practical setting, we use standard analytical message passing rules when available and resort to reparameterization gradient-based message passing (RGMP) only when necessary. In this way, the variance of the free energy gradient, which is due to the sampling process that accompanies RGMP, can be reduced in comparison to the scenario where all messages were based on RGMP. In summary, the contributions of this paper include the following:

- We introduce a new method to approximate posterior distributions in factor graphs by RGMP.
- RGMP can be combined easily with standard update rules such as BP, VMP and EP and extends the class of models for which inference by message passing can be achieved.
- We present hybrid message passing in factor graphs as a natural framework to reduce the variance of the reparameterization gradient estimators.

II. Forney-style Factor Graphs

In this section, we shortly rehearse the Forney-style factor graph framework. Consider the following factorized probabilistic model:

$$\begin{aligned} &p(y,z,x,s,v,w) = \\ &\qquad p(y|z)\cdot p(z|x)\cdot p(x|s)\cdot p(x|v,w)\cdot p(s)\cdot p(v)\cdot p(w)\,. \end{aligned} \tag{2}$$

This model can be graphically represented by the Forney-style factor graph (FGG) in Fig. 1. In an FFG, each node corresponds to a factor and each edge to a variable. An edge connects to a node if the edge variable is an argument of that node’s factor. In Fig. 1, we have associated factors $f_A(v) = p(v)$, $f_B(w) = p(w)$, $f_C(x,v,w) = p(x|v,w)$ etc. with the probability distributions. Since an edge maximally connects to two nodes, most FFGs feature “equality nodes” that can be interpreted as branching nodes. With this accommodation, every factorized probability distribution can be visually represented by an FFG.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

Now assume that variable y is observed (at value $\hat{y}$) and we are interested in inferring the marginal posterior $f(z) \triangleq p(z|y=\hat{y})$. Through marginalization over the latent variables x, s, v, w, we can obtain an expression for the "unnormalized" marginal $\tilde{f}(z)$:

$$\tilde{f}(z) = \underbrace{f_H(\hat{y}, z)}_{m_{f_H \to z}(z)} \cdot \underbrace{\left\{ \int f_G(z,x) \underbrace{\left\{ \int f_E(s) f_D(x,s)\,\mathrm{d}s \right\}}_{m_{f_D \to x}(x)} \cdot \underbrace{\left\{ \iint f_C(x,v,w) f_A(v) f_B(w)\,\mathrm{d}v\mathrm{d}w \right\}}_{m_{f_C \to x}(x)} \mathrm{d}x \right\}}_{m_{f_G \to z}(z)} \quad (3)$$

with exact marginal posterior

$$p(z|y=\hat{y}) = f(z) = \frac{\tilde{f}(z)}{\int \tilde{f}(z)\,\mathrm{d}z}\,. \quad (4)$$

Since each factor only depends on a subset of variables, it is possible to re-distribute the integrands in Eq. 3 such that the entire marginalization process can be interpreted as a message passing algorithm (known as "belief propagation", (BP)), where the calculation of each message only uses locally available information. Also, note that the unnormalized posterior for a variable in BP is obtained by multiplying colliding messages on an edge, i.e., $\tilde{f}(z) = m_{f_H \to z}(z) \cdot m_{f_G \to z}(z)$.

The computation of belief propagation messages and the exact marginal posterior in Eq. 3 and Eq. 4 may not be analytically tractable. For intractable messages, an approximate inference procedure based on variational inference may be applicable. For probabilistic models that are composed of distributions from the exponential family, a message passing procedure known as "variational message passing" (VMP) may lead to analytically tractable message update rules. In particular, in the above example, for the so-called mean-field assumption $q(z,x,s,v,w) = q_z(z)q_x(x)q_s(s)q_v(v)q_w(w)$, the update rule [8] for variable z is given by

$$q_z(z) \propto \exp(\mathbb{E}_{q_x(x)}[\log f_G(z,x) + \log f_H(\hat{y}, z)])\,.$$

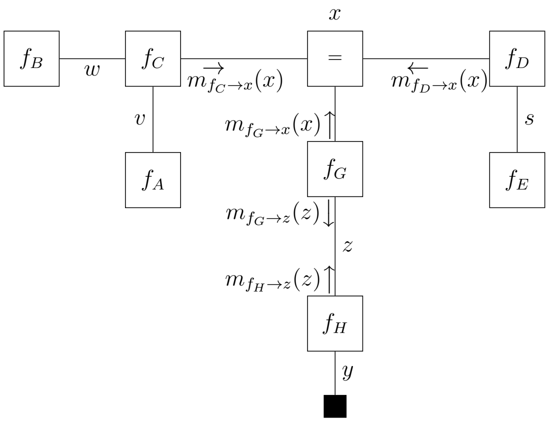
f_B
w
f_C
$\overrightarrow{m}_{f_C \to x}(x)$
x
$=$
$\overleftarrow{m}_{f_D \to x}(x)$
f_D
v
$m_{f_G \to x}(x)\uparrow$
s
f_A
f_G
f_E
$m_{f_G \to z}(z)\downarrow$
z
$m_{f_H \to z}(z)\uparrow$
f_H
y

Fig. 1. A Forney-style factor graph representation of the factorization of model Eq. 2. Nodes represent factors and edges correspond to variables. By convention, edges of observed variables are terminated by a small black node.

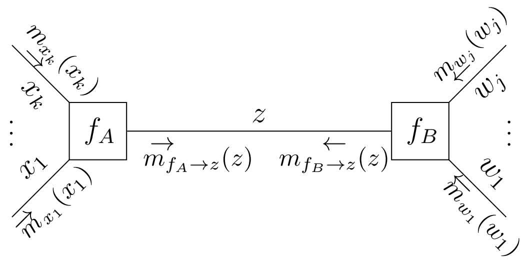

Fig. 2. A generic node-edge-node section in a Forney-style factor graph.

BP and VMP (and other message update rules such as EP) may be combined to perform approximate inference in factor graphs. We refer to [1] and [2] for a more elaborate treatment of message passing-based inference.

III. Problem Formulation

Continuing with the model of Fig. 1, let us assume the following factor specifications:

$$f_G(z,x) = \mathcal{N}(z;\, \alpha x, \nu) \quad (5a)$$

$$f_H(y,z) = \mathcal{P}o(y;\, \exp(z))\,, \quad (5b)$$

where $\mathcal{N}$ and $\mathcal{P}o$ refer to the normal and Poisson distributions. Suppose that $y = \hat{y}$ is observed and in order to proceed with inference, we need to pass messages from variable y up the graph through nodes f_H and f_G. It turns out that for model assumption Eq. 5, the calculations of both the marginal $f(z)$ (per Eq. 4) and the message $m_{f_G \to x}(x)$ have no closed-form solutions in BP and VMP.

In this paper, we develop an alternative variational inference technique for FFGs based on the reparameterization gradient variational inference (RGVI) method [5]–[7]. This technique can be implemented to a large class of continuous distributions and in principle, discrete-valued distribution families are also included if we replace the discrete random variables with their continuous relaxations [9].

IV. Message Passing with the Reparameterization Gradient

In this section, the RGVI method is adapted to message passing in factor graphs. We also discuss how message passing in factor graphs leads to a natural framework for low variance gradient estimators. Finally, we illustrate the reparameterization gradient message passing technique with an example.

A. Approximating the Posterior

Consider a generic node-edge-node section of a FFG in Fig. 2. The edge represents the (vector) random variable $\boldsymbol{z}$, for which we desire to infer the posterior distribution.

Assume that the messages from nodes f_A and f_B to $\boldsymbol{z}$ are computed via BP or VMP, but a difficulty arises in the normalization of $\tilde{f}(\boldsymbol{z})$ (per Eq. 4). In this case, one can approximate the posterior $f(\boldsymbol{z})$ by a variational distribution $q(\boldsymbol{z};\boldsymbol{\phi})$, with variational parameters $\boldsymbol{\phi}$, through minimizing a divergence metric between f and q. Taking the reverse Kullback-Leibler (KL) divergence as the metric, the criterion to be minimized is also known as the variational free energy given in Eq. 1. Choosing the variational distribution family in advance, the task turns into estimating the optimal variational parameters $\boldsymbol{\phi}^*$ that minimizes the free energy function:

$$F(\boldsymbol{\phi}) \triangleq \mathbb{E}_q\left[\log\frac{q(\boldsymbol{z};\boldsymbol{\phi})}{\tilde{f}(\boldsymbol{z})}\right] = \int q(\boldsymbol{z};\boldsymbol{\phi})\log\frac{q(\boldsymbol{z};\boldsymbol{\phi})}{\tilde{f}(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z} \tag{6}$$

where $\tilde{f}(\boldsymbol{z}) = m_{f_A\to z}(\boldsymbol{z})\cdot m_{f_B\to z}(\boldsymbol{z})$.

For a broad range of models, the optimal $\boldsymbol{\phi}^*$ can not be found analytically. Even the gradient $\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi})$ may be hard to evaluate, which hinders iterative optimization. Alternatively, the gradient could be estimated by a Monte Carlo approximation $\hat{\nabla}_{\boldsymbol{\phi}}F(\boldsymbol{\phi}) = \frac{1}{S}\sum_{s=1}^{S}\nabla_{\boldsymbol{\phi}}\log[q(\boldsymbol{z}^{(s)};\boldsymbol{\phi})/\tilde{f}(\boldsymbol{z}^{(s)})]$ where $\{\boldsymbol{z}^{(s)}\}_{s=1}^{S}$ is a set of samples from $q(\boldsymbol{z};\boldsymbol{\phi})$. Unfortunately, the noisy measurements from $\log[q(\boldsymbol{z};\boldsymbol{\phi})/\tilde{f}(\boldsymbol{z})]$ can not be taken without loosing some information about the variational parameters $\boldsymbol{\phi}$. In other words, once $\boldsymbol{z}^{(s)}$ sampled, the term $\log\tilde{f}(\boldsymbol{z}^{(s)})$ is not a function of $\boldsymbol{\phi}$ any more and its gradient $\nabla_{\boldsymbol{\phi}}\log\tilde{f}(\boldsymbol{z}^{(s)})$ becomes zero. RGVI deals with this problem by generating the $\boldsymbol{z}^{(s)}$ samples from a differentiable process of dummy random variables $\boldsymbol{\epsilon}^{(s)}$. Consider a sampling process from a multivariate normal distribution $\mathcal{N}(\boldsymbol{a}, BB^T)$ with mean $\boldsymbol{a}$ and covariance matrix BB^T. In theory, there is no difference between directly sampling from $\mathcal{N}(\boldsymbol{a}, BB^T)$ and generating the samples with the function $g(\boldsymbol{\epsilon};\boldsymbol{a},B) = \boldsymbol{a} + B\cdot\boldsymbol{\epsilon}^{(s)}$ where $\boldsymbol{\epsilon}^{(s)}$ is a sample from $\mathcal{N}(\mathbf{0},\mathbf{I})$. In practice, however, it allows stochastic optimization by tying the samples to the variational parameters. This is a reparameterization example for a normally distributed random variable. More generally, a random variable can be reparameterized as

$$\boldsymbol{\epsilon}^{(s)} \sim p_{\boldsymbol{\epsilon}}(\cdot) \tag{7a}$$

$$\boldsymbol{z}^{(s)} = g(\boldsymbol{\epsilon}^{(s)};\boldsymbol{\phi}) \tag{7b}$$

$$q(\boldsymbol{z};\boldsymbol{\phi}) = \left|\frac{\partial g^{-1}(\boldsymbol{z};\boldsymbol{\phi})}{\partial \boldsymbol{z}}\right| p_{\boldsymbol{\epsilon}}(g^{-1}(\boldsymbol{z};\boldsymbol{\phi})), \tag{7c}$$

where $g(\boldsymbol{\epsilon};\boldsymbol{\phi})$ is an injective, differentiable function of a random variable $\boldsymbol{\epsilon}$, $g^{-1}(\cdot;\boldsymbol{\phi})$ is its inverse, $p_{\boldsymbol{\epsilon}}(\cdot)$ is the probability distribution over ϵ and $\left|\frac{\partial g^{-1}(\boldsymbol{z};\boldsymbol{\phi})}{\partial \boldsymbol{z}}\right|$ is the determinant of the Jacobian for multidimensional $\boldsymbol{\epsilon}$ and $\boldsymbol{z}$, [5]–[7], [10].

The gradient of the free energy can now be estimated by Monte Carlo approximation because it can be expressed as an expectation of the gradient:

$$\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi}) = \mathbb{E}_{p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})}\left[\nabla_{\boldsymbol{\phi}}\log\frac{p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})\left|\frac{\partial\boldsymbol{\epsilon}}{\partial\boldsymbol{z}}\right|}{\tilde{f}(g(\boldsymbol{\epsilon};\boldsymbol{\phi}))}\right]. \tag{8}$$

The above expression is further simplified by discarding the terms that do not include the variational parameters [6], and the result is called the *reparameterization gradient* [5], [7]:

$$\begin{aligned}\nabla_{\boldsymbol{\phi}}F(\boldsymbol{\phi}) &= -\nabla_{\boldsymbol{\phi}}\log\left|\frac{\partial\boldsymbol{z}}{\partial\boldsymbol{\epsilon}}\right| - \mathbb{E}_{p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})}\left[\nabla_{\boldsymbol{\phi}}\log\tilde{f}(g(\boldsymbol{\epsilon};\boldsymbol{\phi}))\right]\\ &= -\nabla_{\boldsymbol{\phi}}\log\left|\frac{\partial\boldsymbol{z}}{\partial\boldsymbol{\epsilon}}\right| - \mathbb{E}_{p_{\boldsymbol{\epsilon}}(\epsilon)}\Big[\nabla_{\phi}\log m_{f_A\to z}\left(g(\boldsymbol{\epsilon};\boldsymbol{\phi})\right)\\ &\qquad + \nabla_{\boldsymbol{\phi}}\log m_{f_B\to z}\left(g(\boldsymbol{\epsilon};\boldsymbol{\phi})\right)\Big].\end{aligned} \tag{9}$$

The variational parameters can now be iteratively updated by employing the gradient estimators $\hat{\nabla}_{\boldsymbol{\phi}}F(\boldsymbol{\phi})$ within a stochastic optimization process such as gradient descent,

$$\boldsymbol{\phi}_{\text{new}} = \boldsymbol{\phi}_{\text{old}} - \rho_i\hat{\nabla}_{\boldsymbol{\phi}}F(\boldsymbol{\phi})|_{\boldsymbol{\phi}=\boldsymbol{\phi}_{\text{old}}}, \tag{10}$$

for varying learning rates ρ_i over the iterations such that the conditions $\sum_{i=1}^{\infty}\rho_i = \infty$, $\sum_{i=1}^{\infty}\rho_i^2 < \infty$ are satisfied [11].

In summary, when closed-form inference is not possible at an edge, we can use a sampling procedure to estimate the "reparameterization gradient" locally at the edge, and use this gradient to minimize the local free energy. In the next section, we use the estimated posterior $q(\boldsymbol{z})$ further to compute outgoing VMP messages.

B. Reparameterization Gradient Message Passing

Continuing with the model of Fig. 1 and specification Eq. 5, the BP message $m_{f_G\to x}(x)$ evaluates to

$$\begin{aligned}m_{f_G\to x}(x) = \int &\left\{\frac{1}{\sqrt{2\pi\nu}}\exp\left(-\frac{(z-\alpha x)^2}{2\nu}\right)\right\}\cdot\\ &\left\{\frac{\exp(z)^{\hat{y}}\exp(-\exp(z))}{\hat{y}!}\right\}\mathrm{d}z\,.\end{aligned} \tag{11}$$

This update rule has no closed-form solution. Fortunately, once the posterior $q(z)$ has been approximated via RGVI, inference can be maintained by variational message passing [2]. We call this procedure *Reparameterization Gradient Message Passing* (RGMP). In our example, we first assume that $f(z)$ is approximated by a Gaussian variational distribution $q(z) = N(z;\mu,\sigma^2)$. Then, unlike for BP and VMP, RGMP allows message passing from f_G to x:

$$\begin{aligned}&m_{f_G\to x}(x) \propto \exp\left(\mathbb{E}_{q(z)}\left[\log f_G(z,x)\right]\right)\\ &\propto \exp\left(\mathbb{E}_{q(z)}\left[-\frac{(z-\alpha x)^2}{2\nu}\right]\right) \propto \mathcal{N}(x;\mu/\alpha,\nu/\alpha^2)\end{aligned} \tag{12}$$

This message may be part of an iterative variational inference process. In practice, it is very common that we can use BP and VMP for all but a few messages in a graph. For those messages, RGMP may then be considered as an alternative method to pass the messages. As discussed, different message passing methods, like BP, VMP and RGMP, may be combined in a factor graph. As a result, RGMP extends the set of models for which we can perform (approximate) inference through message passing.

C. Reducing the Variance of the Gradient Estimators

RGMP makes message passing-based inference possible for a broad range of models. Due to the need for sampling,

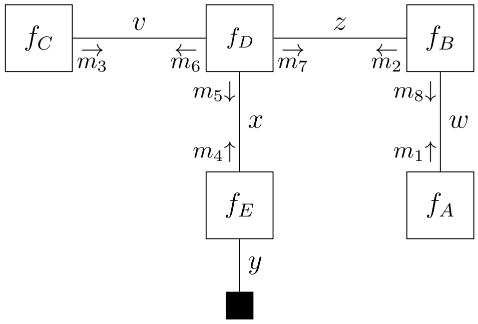

Fig. 3. The example model of Sec. IV-D where y is observed and we are interested in posterior of latent variables x, v, z, w.

reparameterization gradient-based variational inference techniques may suffer from high variance gradients. As a variance reduction technique, sampling processes can be replaced by analytical computations wherever possible. In a factor graph, we can use BP and VMP (with analytical marginalization) at all locations where possible and resort to sampling-based RGMP only when needed. In this way, nuisance variables are analytically marginalized away and these variables do not contribute to the variance in the gradient estimates.

D. Illustrative Example

In Fig. 3, we provide an example for a hybrid message passing-based inference procedure. In this model, assume that y is observed and we are interested in the posteriors of random variables x, z, v, w. Algorithm 1 lists pseudo-code to execute a message passing-based inference algorithm. We assume that the update rules for (m_1, m_2, m_3, m_4) and (m_5, m_8) can be computed analytically by BP and VMP, respectively. However, analytical update rules are not available for m_6 and m_7, so in order to perform inference in this model, we resort to RGMP for those messages. A fully factorized variational posterior $q_x(x)q_z(z)q_v(v)$ can be computed by iterating over a sequence of messages (m_5, m_6, m_7), as indicated in the figure. In this example, the posterior distribution of x is approximated by a stochastic variational procedure as described by [6]. The variational distribution family of $q_z(z)$ is chosen as a Gaussian $\mathcal{N}(z; \mu, \sigma^2)$ with mean parameter μ and variance parameter σ^2 to be estimated. Once the convergence condition is satisfied for the posterior over x, z and v, the marginal posterior of w can be evaluated in one step (see Algorithm 1 for details).

V. Experimental Validation

As an experimental validation of RGMP, we simulated[1] a Poisson Linear Dynamical System (PLDS) [12]. The generative model for this example is given by

$$p(y, x, z) = p(z_0) \prod_{t=1}^{T} p(z_t \mid z_{t-1}) p(x_t \mid z_t) p(y_t \mid x_t) \tag{13}$$

Algorithm 1 Pseudo-code for hybrid message passing-based inference in the model of Fig. 3.

```
Compute $m_1 = m_{f_A \to w}(w)$, $m_2 = m_{f_B \to z}(z)$,
        $m_3 = m_{f_C \to v}(v)$, $m_4 = \log m_{f_E \to x}(x)$ (BP)
Initialize $q_z(z)$, $q_v(v)$, $q_x(x; \mu_{\text{new}}, \sigma^2_{\text{new}})$
repeat
  $m_5 \propto \mathbb{E}_{q_z(z)q_v(v)}[\log f_D(x, z, v)]$ (VMP)
  repeat
    Set learning rates $\rho_\mu$, $\rho_\sigma$ for new iteration
    $\mu_{\text{old}}, \sigma_{\text{old}} = \mu_{\text{new}}, \sigma_{\text{new}}$
    $\epsilon^{(s)} \sim \mathcal{N}(0, 1)$
    $x^{(s)} = \mu_{\text{old}} + \sigma_{\text{old}} \cdot \epsilon^{(s)}$
    $\nabla_\mu^{(s)} F = -\nabla_x \left(m_5(x) + m_4(x)\right)|_{x=x^{(s)}}$
    $\nabla_\sigma^{(s)} F = -\epsilon^{(s)} \nabla_x \left(m_5(x) + m_4(x)\right)|_{x=x^{(s)}} - 1/\sigma_{\text{old}}$
    $\mu_{\text{new}} = \mu_{\text{old}} - \rho_\mu \nabla_\mu^{(s)} F$
    $\sigma_{\text{new}} = \sigma_{\text{old}} - \rho_\sigma \nabla_\sigma^{(s)} F$
  until Convergence
  $q_x(x) = q_x(x; \mu_{\text{new}}, \sigma^2_{\text{new}})$
  $m_6 \propto \exp\left(\mathbb{E}_{q_x(x)q_z(z)}[\log f_D(x, z, v)]\right)$ (RGMP)
  $q_v(v) \propto m_6 \cdot m_3$
  $m_7 \propto \exp\left(\mathbb{E}_{q_x(x)q_v(v)}[\log f_D(x, z, v)]\right)$ (RGMP)
  $q_z(z) \propto m_7 \cdot m_2$
until Convergence
$m_8 \propto \exp\left(\mathbb{E}_{q_z(z)}[\log f_B(z, w)]\right)$ (VMP)
$q_w(w) \propto m_8 \cdot m_1$
```

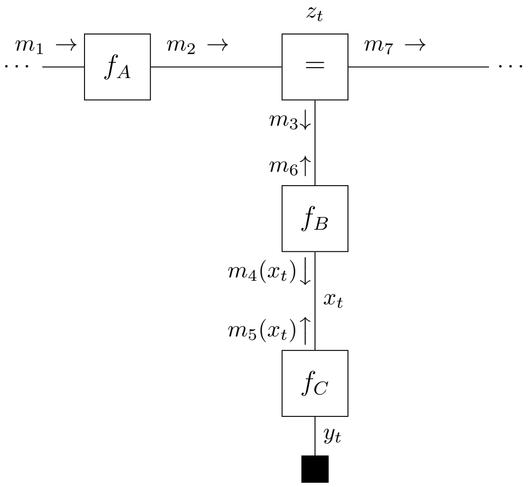

Fig. 4. The factor graph at time slice t of the PLDS model of Sec. V.

where

$$p(z_0) = \mathcal{N}(z_0; \mu^{z_0}, \nu^{z_0}) \tag{14a}$$

$$p(z_t \mid z_{t-1}) = \mathcal{N}(z_t; \alpha z_{t-1}, \nu^z) \tag{14b}$$

$$p(x_t \mid z_t) = \mathcal{N}(x_t; \beta z_t, \nu^x) \tag{14c}$$

$$p(y_t \mid x_t) = \mathcal{P}o(y_t; \exp(x_t)) \,. \tag{14d}$$

One time slice of the factor graph of this model with factors $f_A(z_{t-1}, z_t) = p(z_t|z_{t-1})$, $f_B(x_t, z_t) = p(x_t|z_t)$ and $f_C(y_t, x_t) = p(y_t|x_t)$ is provided in Fig. 4.

[1] Details can be found at https://github.com/biaslab/Semih-EUSIPCO-2019.

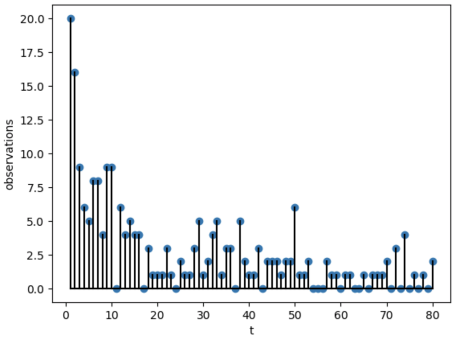

Fig. 5. A synthetic data set of nonnegative integer observations, generated by running model Eq. 14 forward in time.

The PLDS is useful to analyze the trend behind non-negative integer sequences, such as the number of clicks to a video over days. A synthetic dataset, visualized in Fig. 5, is generated for experimental purposes by running model Eq. 14 forward in time with parameter values $\alpha = 0.95$, $\beta = 0.25$, $\nu^z = 0.2$, $\nu^x = 0.1$, $\mu^{z_0} = 12$, $\nu^{z_0} = 0$.

In our simulation, we use a recognition model given by Eq. 14 to model the synthetic time series. In the recognition model, we used a prior $p(z_1) = \mathcal{N}(z_1; 10, 10)$ and noise process parameters $\nu^z = 0.4$ and $\nu^x = 0.5$. The other parameters are the same as for the data generating process. We are interested in recursively updating the posterior for the hidden state z_t from past observations $y_{1:t}$.

Assume that message m_1 carries the variational belief $q(z_{t-1}|y_{1:t-1})$. We are interested in updating to the belief $q(z_t|y_{1:t})$ after a new observation x_t becomes available. This inference task can be executed by message passing schedule $(m_1, m_2, \ldots, m_7)$. However, due to the Poisson likelihood function in PLDS, it is not possible to get a closed-form update rule for inference of x_t, which also blocks computation of later messages such as m_6. To remedy the inference process, we infer a posterior for x_t by applying reparameterization gradient VI to the $m_4(x_t)$ and $m_5(x_t)$ messages. We chose a Gaussian variational distribution $q(x_t) \propto m_4(x_t) \cdot m_5(x_5)$ with

$$\begin{aligned} m_4(x_t) &= \mathcal{N}(x_t; \beta\mu_3, \beta^2\sigma_3^2 + \nu^x) \\ m_5(x_t) &= \mathcal{P}o(y_t; \exp(x_t))\,. \end{aligned}$$

We can use posterior $q(x_t)$ to update message m_6 by

$$\begin{aligned} m_6(z_t) &\propto \exp\left(\mathbb{E}_{q(x_t)}[\log f_B(x_t, z_t)]\right) \\ &\propto \mathcal{N}(z_t; \langle x_t \rangle_{q(x_t)} / \beta, \nu^x/\beta^2)\,. \end{aligned}$$

Message m_6 can then be used within a standard message passing method for Linear Gaussian Dynamical Systems [1].

The mean and standard deviation for the estimated posterior over z_t is visualized in Fig. 6. The "true" value for z_t lies almost everywhere within the one-standard-deviation range.

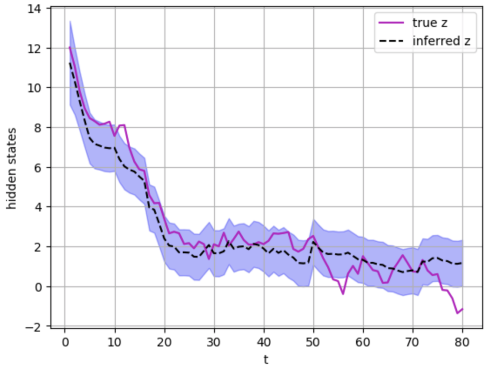

Fig. 6. Estimated mean and standard deviation over time for the state dynamics z_t of model Eq. 14.

VI. Conclusions

In this paper, we introduced reparameterization gradient message passing, which is an adaptation of the reparameterization gradient variational inference method to factor graphs. Since RGMP can easily be combined with BP and VMP, the proposed method extends the reach of message passing-based inference methods beyond existing techniques. While definitive performance results are scheduled for future work, we presented promising results on a dynamic state estimation task with a Poisson likelihood function.

References

[1] 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, June 2007.

[2] John Winn and Christopher M Bishop. Variational Message Passing. *Journal of Machine Learning Research*, 6(Apr):661–694, April 2005.

[3] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. arXiv: 1601.00670.

[4] Thomas P. Minka. Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[5] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, December 2013. arXiv: 1312.6114.

[6] Michalis K. Titsias and Miguel Lzaro-Gredilla. Doubly Stochastic Variational Bayes for Non-conjugate Inference. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1971–II–1980, Beijing, China, 2014. JMLR.org.

[7] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv:1401.4082 [cs, stat]*, January 2014. arXiv: 1401.4082.

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

[9] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *arXiv:1611.00712 [cs, stat]*, November 2016. arXiv: 1611.00712.

[10] Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. *arXiv:1505.05770 [cs, stat]*, May 2015. arXiv: 1505.05770.

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

[12] Evan Archer, Il Memming Park, Lars Buesing, John Cunningham, and Liam Paninski. Black box variational inference for state space models. *arXiv:1511.07367 [stat]*, November 2015. arXiv: 1511.07367.