Variational inference in (sparse) latent LDS models

bcseke@inf.ed.ac.uk

January 17, 2013

Abstract

Factored expectation constraints based approximate inference for Latent LDS models. Details of sparse models are also considered.

Model

We consider the stationary homogeneous latent Gaussian model

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t + \boldsymbol{Q}^{-1/2}\boldsymbol{w}_t, \quad \text{with} \quad \boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \\ p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}) &= \phi_{t+1,j}(\boldsymbol{x}_{t+1}; \boldsymbol{y}_{t+1}) \end{aligned}$$

where the inputs $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T]$ and the outputs $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$ are observed and known and $\phi_{t+1,j}$ is some, possibly non-Gaussian observation model. We put independent priors on the elements of the parameters $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{Q}$ and write the joint density conditioned by the inputs as

$$p(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{B}|\boldsymbol{U}) = \left[\prod_{ij} p_0(a_{ij}) \prod_{ij} p_0(b_{ij}) \prod_{ij} p_0(q_{ij})\right] \times \prod_t N(\boldsymbol{x}_{t+1}|\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{Q}^{-1}) \prod_j \phi_{t+1,j}(\boldsymbol{x}_{t+1}; \boldsymbol{y}_{t+1}).$$

In the following we show how we plan to do inference with a factored expectation propagation algorithm that is expected to get closer to the free form variational inference than the fixed form variational approach. We start from the free form variational approach and then we present the details of the expectation constraints based procedure.

Free form variational inference

Here we choose to approximate the joint density $p(\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{B}|\boldsymbol{Y}, \boldsymbol{U})$ with a factored form $q_x(\boldsymbol{X})q_A(\boldsymbol{A})q_Q(\boldsymbol{Q})q_B(\boldsymbol{B})$ by using the KL divergence $\mathrm{D}[\cdot||\cdot]$, that is we have

$$\underset{q_x, q_A, q_Q, q_B}{\text{minimise}} \; \mathrm{D}[q_x(\boldsymbol{X})q_A(\boldsymbol{A})q_Q(\boldsymbol{Q})q_B(\boldsymbol{B})||p(\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{B}|\boldsymbol{Y}, \boldsymbol{U})] \,.$$

The stationary point of this minimisation problem yield the structured variational mean field updates

$$[q_x(\boldsymbol{X})]^{new} \propto \exp\left\{\langle \log p(\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{B}|\boldsymbol{Y}, \boldsymbol{U})\rangle_{q_A q_Q, q_B}\right\} \tag{1}$$

$$[q_A(\boldsymbol{A})]^{new} \propto \exp\left\{\langle \log p(\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{B}|\boldsymbol{Y}, \boldsymbol{U})\rangle_{q_X q_Q, q_B}\right\} \tag{2}$$

$$[q_Q(\boldsymbol{Q})]^{new} \propto \exp\left\{\langle \log p(\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{B}|\boldsymbol{Y}, \boldsymbol{U})\rangle_{q_X q_A, q_B}\right\} \tag{3}$$

$$[q_B(\boldsymbol{B})]^{new} \propto \exp\left\{\langle \log p(\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{B}|\boldsymbol{Y}, \boldsymbol{U})\rangle_{q_X q_A, q_Q}\right\} \tag{4}$$

From now on we will use and alternative notation for expectations: and expectation $h_i(x_i) \equiv \langle h(x_1, \ldots, x_i, \ldots, x_n)\rangle_{q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_n}$ is also denoted as $h_i(x_i) = \langle h(x_1, \ldots, x_i, \ldots, x_n)\rangle_{\setminus q_i}$. It

will typically be clear from the context what the densities $q_1, \ldots, q_n$ are and which member of this set of densities is omitted.

In the following sections we detail the form of all these densities and models in turn and evaluate what further steps can be taken to approximate them. The crucial term is the transition term that can be written as

$$\log N(\boldsymbol{x}_{t+1}|\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{Q}^{-1}) = \frac{1}{2}\log\det(\boldsymbol{Q}/2\pi) - \frac{1}{2}(\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{B}\boldsymbol{u}_t)^T\boldsymbol{Q}(\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{B}\boldsymbol{u}_t) \quad (5)$$

As we can see, the above separation of the variables is really justified: due to the form of the interaction there joint modelling of the state space and the parameters is problematic as there is hardly any parametric class that could deal with such high degree of interaction, whereas when considered conditionally, they are all quadratic.

The model for q_x

From (5) we obtain a q_x that can be rewritten as

$$q_x(\boldsymbol{X}) \propto \prod_t \Psi_{t,t+1}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) \prod_j \phi_{t+1,j}(\boldsymbol{x}_{t+1}; \boldsymbol{y}_{t+1}) \quad (6)$$

with $\Psi_{t,t+1}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t) = \left\langle \log N(\boldsymbol{x}_{t+1}|\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{Q}^{-1}) \right\rangle_{q_A q_B q_Q}$ is a canonical Gaussian with parameters $\left\langle \boldsymbol{h}^x_{t,t+1} \right\rangle_{\setminus q_x}$ and $\left\langle \boldsymbol{Q}^x_{t,t+1} \right\rangle_{\setminus q_x}$ where

$$\boldsymbol{h}^x_{t,t+1} = \begin{bmatrix} -\boldsymbol{A}^T\boldsymbol{Q}\boldsymbol{B}\boldsymbol{u}_t \\ \boldsymbol{0} \end{bmatrix}$$
$$\boldsymbol{Q}^x_{t,t+1} = \begin{bmatrix} \boldsymbol{A}^T\boldsymbol{Q}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{Q} \\ -\boldsymbol{Q}\boldsymbol{A} & \boldsymbol{Q} \end{bmatrix}.$$

Therefore, we have a latent Gaussian model and inference in this model can be addressed by well-known approximations.

The model for q_A

Let is use the notation $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$ and let $[\boldsymbol{A}]_c$ be the column ordered vectorised form of $\boldsymbol{A}$, namely $[\boldsymbol{A}]_c^T = [\boldsymbol{a}_1^T, \ldots, \boldsymbol{a}_n^T]$. Then it follows that

$$q_A(\boldsymbol{A}) \propto \exp\left\{ \langle \boldsymbol{h}_A \rangle^T_{\setminus q_A} [\boldsymbol{A}]_c - \frac{1}{2}[\boldsymbol{A}]_c^T \langle \boldsymbol{Q}_A \rangle_{\setminus q_A} [\boldsymbol{A}]_c \right\} \times \prod_{ij} p_0(a_{ij}), \quad (7)$$

where

$$\boldsymbol{h}_A = \left[\boldsymbol{Q} \sum_t \boldsymbol{x}_{t+1}\boldsymbol{x}_t^T - \boldsymbol{Q}\boldsymbol{B} \sum_t \boldsymbol{u}_t\boldsymbol{x}_t^T \right]_c,$$
$$\boldsymbol{Q}_A = \sum_t \boldsymbol{x}_t\boldsymbol{x}_t^T \otimes \boldsymbol{Q}$$

The model for q_B

The model for $\boldsymbol{B}$ follows a similar structure and is given by

$$q_B(\boldsymbol{B}) \propto \exp\left\{ \langle \boldsymbol{h}_B \rangle^T_{\setminus q_B} [\boldsymbol{B}]_c - \frac{1}{2}[\boldsymbol{B}]_c^T \langle \boldsymbol{Q}_B \rangle_{\setminus q_B} [\boldsymbol{B}]_c \right\} \times \prod_{ij} p_0(b_{ij}) \quad (8)$$

with

$$\boldsymbol{h}_B = \left[\boldsymbol{Q}\sum_t (\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t)\boldsymbol{u}_t^T\right]_c,$$
$$\boldsymbol{Q}_B = \sum_t \boldsymbol{u}_t\boldsymbol{u}_t \otimes \boldsymbol{Q}.$$

The model for q_Q

This model is rather complicated because it involves a log determinant term, but suitable choices for the structure of $\boldsymbol{Q}$ and the prior $p_0(\boldsymbol{Q})$ can help to make inference tractable. The distribution q_Q has the form

$$q_Q(\boldsymbol{Q}) \propto \exp\left\{\frac{1}{2}\log\det\boldsymbol{Q} - \frac{1}{2}\operatorname{tr}(\boldsymbol{Q}^T\boldsymbol{H}_Q)\right\} \times \prod_{ij} p_0(q_{ij}), \tag{9}$$

where

$$\begin{aligned}\boldsymbol{H}_q = \langle \boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}\rangle_{q_x} - 2\operatorname{tr}\left\{\langle\boldsymbol{A}\rangle_{q_A}\left\langle\boldsymbol{x}_t\boldsymbol{x}_{t+1}^T\right\rangle_{q_x}\right\} + \operatorname{tr}\left\{\left\langle\boldsymbol{A}^T\boldsymbol{A}\right\rangle_{q_A}\left\langle\boldsymbol{x}_t\boldsymbol{x}_t^T\right\rangle_{q_x}\right\} \\ - 2\left\langle\boldsymbol{B}\right\rangle_{q_B}\sum_t \boldsymbol{u}_t(\langle\boldsymbol{x}_{t+1}\rangle_{q_x} - \langle\boldsymbol{A}\rangle_{q_A}\langle\boldsymbol{x}_t\rangle_{q_x})^T + \left\langle\boldsymbol{B}^T\boldsymbol{B}\right\rangle_{q_B}\sum_t \boldsymbol{u}_t\boldsymbol{u}_t^T.\end{aligned}$$

Inference with factored expectation propagation

With the exception of q_Q, all model above are latent Gaussian models where exact inference can be intractable due to the priors or the likelihoods. For this reason we will introduce a factored EP algorithm which basically simplifies to doing EP in these models. In the following we introduce a family of marginals, expectation constraints and the corresponding entropy approximation for each of the above models.

Approximate marginals for q_x

We define the family of approximate marginal densities

$$\mathcal{Q}_x = \{\{q^x_{t,t+1}\}_t, \{q^{x,f}_t\}_t, \{q^{x,g}_{t,j}\}_{t,j}, \{q^g_{t,j}\}_{t,j}\}. \tag{10}$$

where we assign the density $q^x_{t,t+1}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$ to $\Psi_{t,t+1}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})$ and $q^x_{t+1,j}(\boldsymbol{x}_{t+1})$ to $\phi_{t+1,j}(\boldsymbol{x}_{t+1})$. This family can be viewed as a set of marginals that defines a joint density

$$q^x(\boldsymbol{X}) \propto \frac{\prod_t q^x_{t,t+1}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1})}{\prod_t q^{x,f}_{t+1}(\boldsymbol{x}_{t+1})} \times \prod_{t,j}\frac{q^x_{t+1,j}(\boldsymbol{x}_{t+1})}{q^{x,l}_{t+1,j}(\boldsymbol{x}_{t+1})}.$$

As we will see later, this representation of q_x is exact up to the normalisation constant, however, the factors themselves are only approximations of the marginal densities. and we assume that expectation constraints w.r.t. $\boldsymbol{f}(\boldsymbol{x}_{t+1})$ and $\boldsymbol{l}_j(\boldsymbol{x}_{t+1})$ hold between $q^x_{t,t+1}$ and $q^x_{t+1,t+2}$ and $q^x_{t,t+1}$ and $q^x_{t+1,j}$ respectively. The families $\boldsymbol{f}$ and $\boldsymbol{l}_j$ will be restricted Gaussians, that is, $\boldsymbol{f}$ and $\boldsymbol{l}_j$ will correspond to a sparse Gaussian Markov random fields. In many cases we consider $\phi_{t+1,j}$s that depend on a subset I_j of $\boldsymbol{x}_{t+1}$s elements or only one element x^j_{t+1}, that is, we have $\phi_{t+1,j}(\boldsymbol{x}_{t+1}; \boldsymbol{y}_{t+1}) = \phi_{t+1,j}(x^{I_j}_{t+1}; \boldsymbol{y}_{t+1})$ or $\phi_{t+1,j}(\boldsymbol{x}_{t+1}; \boldsymbol{y}_{t+1}) = \phi_{t+1,j}(x^j_{t+1}; y^j_{t+1})$ and thus $\boldsymbol{l}_j(\boldsymbol{x}_{t+1}) = \boldsymbol{l}_j(\boldsymbol{x}^{I_j}_{t+1})$ or $\boldsymbol{l}_j(\boldsymbol{x}_{t+1}) = \boldsymbol{l}_j(x^j_{t+1})$. The choice of these sufficient statistics depends on the specific model at hand and, as we will see later, our selection criteria will be the tractability of the computations they require, that is, the sparsity they generate.

Since we ...

The expectation constraints can be written as

$$\langle \boldsymbol{f}(\boldsymbol{x}_{t+1})\rangle_{q^x_{t,t+1}} = \langle \boldsymbol{f}(\boldsymbol{x}_{t+1})\rangle_{q^{x,f}_{t+1}} \quad \text{and} \quad \langle \boldsymbol{f}(\boldsymbol{x}_{t+1})\rangle_{q^x_{t+1,t+2}} = \langle \boldsymbol{f}(\boldsymbol{x}_{t+1})\rangle_{q^{x,f}_{t+1}}\,, \tag{11}$$

and

$$\langle \boldsymbol{l}_j(\boldsymbol{x}_{t+1})\rangle_{q^x_{t,t+1}} = \langle \boldsymbol{l}_j(\boldsymbol{x}_{t+1})\rangle_{q^{x,l}_{t+1,j}} \quad \text{and} \quad \langle \boldsymbol{l}_j(\boldsymbol{x}_{t+1})\rangle_{q^x_{t+1,j}} = \langle \boldsymbol{l}_j(\boldsymbol{x}_{t+1})\rangle_{q^{x,l}_{t+1,j}}\,, \tag{12}$$

respectively.

As a consequence we introduce the approximate entropy

$$-\tilde{H}(\mathcal{Q}_x) = \sum_t \left\langle \log q^x_{t,t+1}\right\rangle_{q^x_{t,t+1}} - \sum_t \left\langle \log q^{x,f}_t\right\rangle_{q^{x,f}_t} + \sum_{t,j} \left\langle \log q^x_{t,j}\right\rangle_{q^x_{t,j}} - \sum_{t,j} \left\langle \log q^{x,l}_{t,j}\right\rangle_{q^{x,l}_{t,j}} \tag{13}$$

and Lagrange multiplier terms

$$\begin{aligned} C(\mathcal{Q}_x, \Lambda_x) = &\sum_t \boldsymbol{\lambda}^{\boldsymbol{\beta}}_{\boldsymbol{t+1}} \cdot \left[\langle \boldsymbol{f}(\boldsymbol{x}_{t+1})\rangle_{q^{x,f}_{t+1}} - \langle \boldsymbol{f}(\boldsymbol{x}_{t+1})\rangle_{q^x_{t,t+1}}\right] + \sum_t \boldsymbol{\lambda}^{\boldsymbol{\alpha}}_{\boldsymbol{t+1}} \cdot \left[\langle \boldsymbol{f}(\boldsymbol{x}_{t+1})\rangle_{q^{x,f}_{t+1}} - \langle \boldsymbol{f}(\boldsymbol{x}_{t+1})\rangle_{q^x_{t+1,t+2}}\right] \\ &\sum_{t,j} \boldsymbol{\lambda}^0_{t+1,j} \cdot \left[\langle \boldsymbol{l}(\boldsymbol{x}_{t+1})\rangle_{q^{x,l}_{t+1,j}} - \langle \boldsymbol{l}(\boldsymbol{x}_{t+1})\rangle_{q^x_{t,t+1}}\right] + \sum_{t,j} \boldsymbol{\lambda}^l_{t+1,j} \cdot \left[\langle \boldsymbol{l}(\boldsymbol{x}_{t+1})\rangle_{q^{x,l}_{t+1,j}} - \langle \boldsymbol{l}(\boldsymbol{x}_{t+1})\rangle_{q^x_{t+1,j}}\right] \end{aligned} \tag{14}$$

Approximate marginals for q_A

In case of q_A we define the family

$$\mathcal{Q}_A = \{q^A_0, \{q^A_{ij}\}_{ij}, \{q^{A,g}_{ij}\}_{ij}\} \tag{15}$$

where we associate q^A_0 with the exponentiated quadratic form in q_A and we associate $q^A_{ij}(a_{ij})$ with the prior $p_0(a_{ij})$. We also define a set of expectation constraints between the members of $\mathcal{Q}_A$. Because of latent Gaussian nature of q_A, we will require expectation constraints up to second order. That is, we choose $\boldsymbol{g}(z) = (z, -z^2/2)$ and we set the constraints $\langle \boldsymbol{g}(a_{ij})\rangle_{q^{A,g}_{ij}} = \langle \boldsymbol{g}(a_{ij})\rangle_{q^A_0}$ and $\langle \boldsymbol{g}(a_{ij})\rangle_{q^A_{ij}} = \langle \boldsymbol{g}(a_{ij})\rangle_{q^A_0}$. The family $\mathcal{Q}_A$ defines a set of marginal densities that corresponds to a density written in the form

$$q^A(\boldsymbol{A}) \propto q^A_0(\boldsymbol{A}) \prod_{ij} \frac{q^A_{ij}(a_{ij})}{q^{A,g}_{ij}(a_{ij})}. \tag{16}$$

The corresponding entropy approximation will be defined as

$$-\tilde{H}(\mathcal{Q}_A) = \left\langle \log q^A_0\right\rangle_{q^A_0} + \sum_{ij} \left\langle \log q^A_{ij}\right\rangle_{q^A_{ij}} - \sum_{ij} \left\langle \log q^{A,g}_{ij}\right\rangle_{q^{A,g}_{ij}}\,, \tag{17}$$

and a set of Lagrangian term for the expectation constraints can be written as

$$C(\mathcal{Q}_A, \Lambda_A) = \sum_{ij} \boldsymbol{\lambda}^A_{g,ij} \cdot \left[\langle \boldsymbol{g}(a_{ij})\rangle_{q^{A,g}_{ij}} - \langle \boldsymbol{g}(a_{ij})\rangle_{q^A_{ij}}\right] + \sum_{ij} \boldsymbol{\lambda}^A_{0,ij} \cdot \left[\langle \boldsymbol{g}(a_{ij})\rangle_{q^{A,g}_{ij}} - \langle \boldsymbol{g}(a_{ij})\rangle_{q^A_0}\right]. \tag{18}$$

Issues related to the sparsity constraint on $\boldsymbol{A}$ will be discussed in Section (?).

Approximate marginals for q_B

In case of q_B we define the family

$$\mathcal{Q}_B = \{q^B_0, \{q^B_{ij}\}_{ij}, \{q^{B,g}_{ij}\}_{ij}\} \tag{19}$$

and we assign q_0^B with the exponentiated quadratic form in q_B and we assign $q_{ij}^B(b_{ij})$ with the prior $p_0(b_{ij})$. We define the expectation constraints similarly as in the section above. The corresponding approximating density can be viewed as

$$q^B(\boldsymbol{B}) \propto q_0^B(\boldsymbol{B}) \prod_{ij} \frac{q_{ij}^B(b_{ij})}{q_{ij}^{B,g}(b_{ij})}. \tag{20}$$

The approximate entropy term will be

$$-\tilde{H}(\mathcal{Q}_B) = \left\langle \log q_0^B \right\rangle_{q_0^B} + \sum_{ij} \left\langle \log q_{ij}^B \right\rangle_{q_{ij}^B} - \sum_{ij} \left\langle \log q_{ij}^{B,g} \right\rangle_{q_{ij}^{B,g}} \tag{21}$$

and the Langrangian term corresponding to the expectation constraints will be written as

$$C(\mathcal{Q}_B, \Lambda_B) = \sum_{ij} \boldsymbol{\lambda}_{g,ij}^B \cdot \left[\langle \boldsymbol{g}(b_{ij}) \rangle_{q_{ij}^{B,g}} - \langle \boldsymbol{g}(b_{ij}) \rangle_{q_{ij}^B} \right] + \sum_{ij} \boldsymbol{\lambda}_{0,ij}^B \cdot \left[\langle \boldsymbol{g}(b_{ij}) \rangle_{q_{ij}^{B,g}} - \langle \boldsymbol{g}(b_{ij}) \rangle_{q_0^B} \right]. \tag{22}$$

Approximate marginals for q_Q

We do not define any approximation for q_Q, we assume that the free form variational update can be done analytically and all the expectation and the entropy term needed are also tractable.

Free energy optimisation and message passing

The Lagrangian corresponding to the free energy minimisation expressed in therm of $\mathcal{Q}_x, \mathcal{Q}_A, \mathcal{Q}_B$ and q_Q is

$$\begin{aligned} L(\mathcal{Q}_x, \mathcal{Q}_A, \mathcal{Q}_Q, \mathcal{Q}_B, \Lambda_x, \Lambda_A, \Lambda_Q, \Lambda_B) = &- \langle \log p(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{B} | \boldsymbol{U}) \rangle_{\mathcal{Q}_x, \mathcal{Q}_A, \mathcal{Q}_Q, \mathcal{Q}_B} \\ &- \tilde{H}(\mathcal{Q}_x) - \tilde{H}(\mathcal{Q}_A) - H(q_Q) - \tilde{H}(\mathcal{Q}_B) \\ &+ C(\mathcal{Q}_x, \Lambda_x) + C(\mathcal{Q}_A, \Lambda_A) + C(\mathcal{Q}_B, \Lambda_B) + \text{normalisation constraints}. \end{aligned}$$

In addition, let us define $\text{Collapse}(p(\boldsymbol{z}); \boldsymbol{f})$ as the (moment) projection of $p(\boldsymbol{x})$ into the exponential family defined by $\boldsymbol{f}(\boldsymbol{x})$, that is,

$$\text{Collapse}[p(\boldsymbol{z}); \boldsymbol{f}] = \underset{\theta}{\text{argmin}}\, \text{D}[p(\boldsymbol{z}) || \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{z}) - \log Z(\boldsymbol{\theta}))]\,.$$

To differentiate between the free from densities, for example q_A, and the corresponding family of marginals, for example $\mathcal{Q}_A$, we use a different kind of notation for expectations. The quantities corresponding to $\boldsymbol{h_A}$ and $\boldsymbol{Q}_A$ will be denoted by $\langle \boldsymbol{h}_A \rangle_{\setminus \mathcal{Q}_A}$ and $\langle \boldsymbol{Q}_A \rangle_{\setminus \mathcal{Q}_A}$ and the same rule applies for q_B and $\mathcal{Q}_B$. For q_x and $\mathcal{Q}_x$ we use the shortcut notation $\langle \log \Psi_{t,t+1}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) \rangle_{\setminus \mathcal{Q}_x}$, but the same usage applies.

From the stationary conditions of L w.r.t the members of the families $\mathcal{Q}_x, \mathcal{Q}_A$ and $\mathcal{Q}_B$ we can derive the form of the approximating densities. The stationarity conditions corresponding to the expectation constraints will be used to define a message passing algorithm.

These are as follows

(1) for the members of the family $\mathcal{Q}_x$ we have

$$q_{t,t+1}^x(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) \propto \exp \Big\{ \langle \log \Psi_{t,t+1}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}) \rangle_{\setminus \mathcal{Q}_x} \tag{23}$$

$$+ \boldsymbol{\lambda}_t^\alpha \cdot \boldsymbol{f}(\boldsymbol{x}_t) + \boldsymbol{\lambda}_{t+1}^\beta \cdot \boldsymbol{f}(\boldsymbol{x}_{t+1}) + \sum_j \boldsymbol{\lambda}_{t+1,j}^0 \cdot \boldsymbol{l}_j(\boldsymbol{x}_{t+1}) \Big\}, \tag{24}$$

$$q_{t+1,j}^x(\boldsymbol{x}_{t+1}) \propto \psi_{t+1,j}(\boldsymbol{x}_{t+1}; \boldsymbol{y}_{t+1}) \times \exp\{\boldsymbol{\lambda}_{t+1,j}^l \cdot \boldsymbol{l}_j(\boldsymbol{x}_{t+1})\}, \tag{25}$$

$$q_{t+1,j}^{x,l}(\boldsymbol{x}_{t+1}) \propto \exp\{(\boldsymbol{\lambda}_{t+1,j}^0 + \boldsymbol{\lambda}_{t+1,j}^l) \cdot \boldsymbol{l}_j(\boldsymbol{x}_{t+1})\}, \tag{26}$$

$$q_{t+1}^{x,f}(\boldsymbol{x}_{t+1}) \propto \exp\{(\boldsymbol{\lambda}_{t+1}^\alpha + \boldsymbol{\lambda}_{t+1}^f) \cdot \boldsymbol{f}(\boldsymbol{x}_{t+1})\}, \tag{27}$$

and the corresponding expectation constraints lead the update equations

$$[\boldsymbol{\lambda}^{l}_{t+1,j}]^{new} = \text{Collapse}(q^{x}_{t,t+1}[\boldsymbol{x}_{t+1}); \boldsymbol{l}_j] - \boldsymbol{\lambda}^{0}_{t+1,j}, \tag{28}$$

$$[\boldsymbol{\lambda}^{0}_{t+1,j}]^{new} = \text{Collapse}(q^{x}_{t+1,j}[\boldsymbol{x}_{t+1}); \boldsymbol{l}_j] - \boldsymbol{\lambda}^{l}_{t+1,j}, \tag{29}$$

$$[\boldsymbol{\lambda}^{\alpha}_{t+1}]^{new} = \text{Collapse}(q^{x}_{t,t+1}[\boldsymbol{x}_{t+1}); \boldsymbol{f}] - \boldsymbol{\lambda}^{\beta}_{t+1}, \tag{30}$$

$$[\boldsymbol{\beta}^{\alpha}_{t+1}]^{new} = \text{Collapse}(q^{x}_{t+1,t+2}[\boldsymbol{x}_{t+1}); \boldsymbol{f}] - \boldsymbol{\lambda}^{\alpha}_{t+1}, \tag{31}$$

(2) for the members of the family $\mathcal{Q}_A$ we have

$$q^{A}_{0} \propto \exp\Big\{[\boldsymbol{A}]^{T}_{c}\,\langle \boldsymbol{h}_A\rangle_{\backslash\mathcal{Q}_A} - \frac{1}{2}[\boldsymbol{A}]^{T}_{c}\,\langle \boldsymbol{Q}_A\rangle_{\backslash\mathcal{Q}_A}\,[\boldsymbol{A}]_c + \sum_{ij}\boldsymbol{\lambda}^{A}_{0,ij}\cdot\boldsymbol{g}(a_{ij})\Big\}, \tag{32}$$

$$q^{A}_{ij} \propto p_0(a_{ij}) \times \exp\left\{\boldsymbol{\lambda}^{A}_{g,ij}\cdot\boldsymbol{g}(a_{ij})\right\}, \tag{33}$$

$$q^{A,g}_{ij} \propto \exp\left\{(\boldsymbol{\lambda}^{A}_{0,ij} + \boldsymbol{\lambda}^{A}_{g,ij})\cdot\boldsymbol{g}(a_{ij})\right\}, \tag{34}$$

and the corresponding expectation constraints lead the equations

$$[\boldsymbol{\lambda}^{A}_{0,ij}]^{new} = \text{Collapse}[q^{A}_{ij}(a_{ij}); \boldsymbol{g}] - \boldsymbol{\lambda}^{A}_{g,ij}, \tag{35}$$

$$[\boldsymbol{\lambda}^{A}_{g,ij}]^{new} = \text{Collapse}[q^{A}_{0}(a_{ij}); \boldsymbol{g}] - \boldsymbol{\lambda}^{A}_{0,ij}, \tag{36}$$

(3) for the members of the family $\mathcal{Q}_B$ we have

$$q^{B}_{0} \propto \exp\Big\{[\boldsymbol{B}]^{T}_{c}\,\langle \boldsymbol{h}_B\rangle_{\backslash\mathcal{Q}_B} - \frac{1}{2}[\boldsymbol{B}]^{T}_{c}\,\langle \boldsymbol{Q}_B\rangle_{\backslash\mathcal{Q}_B}\,[\boldsymbol{B}]_c + \sum_{ij}\boldsymbol{\lambda}^{B}_{0,ij}\cdot\boldsymbol{g}(b_{ij})\Big\}, \tag{37}$$

$$q^{B}_{ij} \propto p_0(b_{ij}) \times \exp\left\{\boldsymbol{\lambda}^{B}_{g,ij}\cdot\boldsymbol{g}(b_{ij})\right\}, \tag{38}$$

$$q^{B,g}_{ij} \propto \exp\left\{(\boldsymbol{\lambda}^{B}_{0,ij} + \boldsymbol{\lambda}^{B}_{g,ij})\cdot\boldsymbol{g}(b_{ij})\right\}, \tag{39}$$

and the corresponding expectation constraints lead the equations

$$[\boldsymbol{\lambda}^{B}_{0,ij}]^{new} = \text{Collapse}[q^{B}_{ij}(a_{ij}); \boldsymbol{g}] - \boldsymbol{\lambda}^{B}_{g,ij}, \tag{40}$$

$$[\boldsymbol{\lambda}^{B}_{g,ij}]^{new} = \text{Collapse}[q^{B}_{0}(a_{ij}); \boldsymbol{g}] - \boldsymbol{\lambda}^{B}_{0,ij}. \tag{41}$$

Details of the collapse operations

The collapse operations on $q^{x}_{t,t+1}$

Let $\boldsymbol{\lambda}^{\alpha}_{t} = (\boldsymbol{h}^{\alpha}_{t}, \boldsymbol{Q}^{\alpha}_{t})$, $\boldsymbol{\lambda}^{\beta}_{t} = (\boldsymbol{h}^{\beta}_{t}, \boldsymbol{Q}^{\beta}_{t})$ and $\boldsymbol{\lambda}^{0}_{t+1,j} = (\boldsymbol{h}^{0}_{t+1,j}, \boldsymbol{Q}^{0}_{t+1,j})$. Then the density $q^{x}_{t,t+1}$ is a multivariate Gaussian with canonical parameters

$$\tilde{\boldsymbol{h}}^{x}_{t,t+1} = \left\langle \boldsymbol{h}^{x}_{t,t+1}\right\rangle_{\backslash\mathcal{Q}_x} + \begin{bmatrix} \boldsymbol{h}^{\alpha}_{t+1,j} \\ \boldsymbol{h}^{\beta}_{t+1,j} + \sum\limits_{j}\boldsymbol{h}^{0}_{t+1,j} \end{bmatrix}, \tag{42}$$

$$= \begin{bmatrix} -\langle \boldsymbol{A}\rangle^{T}_{\mathcal{Q}_A}\langle \boldsymbol{Q}\rangle_{q_Q}\langle \boldsymbol{B}\rangle_{\mathcal{Q}_B}\boldsymbol{u}_t + \boldsymbol{h}^{\alpha}_{t+1,j} \\ \boldsymbol{h}^{\beta}_{t+1,j} + \sum\limits_{j}\boldsymbol{h}^{0}_{t+1,j} \end{bmatrix}, \tag{43}$$

$$\tilde{Q}^{x}_{t,t+1} = \left\langle \boldsymbol{Q}^{x}_{t,t+1}\right\rangle_{\backslash\mathcal{Q}_x} + \begin{bmatrix} \boldsymbol{Q}^{\alpha}_{t} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Q}^{\beta}_{t+1} + \sum\limits_{j}\boldsymbol{Q}^{0}_{t+1,j} \end{bmatrix} \tag{44}$$

$$= \begin{bmatrix} \left\langle \boldsymbol{A}^{T}\langle \boldsymbol{Q}\rangle_{q_Q}\boldsymbol{A}\right\rangle_{\mathcal{Q}_A} + \boldsymbol{Q}^{\alpha}_{t} & -\langle \boldsymbol{A}\rangle^{T}_{\mathcal{Q}_A}\langle \boldsymbol{Q}\rangle_{q_Q} \\ -\langle \boldsymbol{Q}\rangle_{q_Q}\langle \boldsymbol{A}\rangle_{\mathcal{Q}_A} & \langle \boldsymbol{Q}\rangle_{q_Q} + \boldsymbol{Q}^{\beta}_{t+1} + \sum\limits_{j}\boldsymbol{Q}^{0}_{t+1,j} \end{bmatrix} \tag{45}$$

and the collapse operations are equivalent to computing the mean and covariance values corresponding to the non-zeros in $\boldsymbol{Q}_t^{\alpha}$, $\boldsymbol{Q}_t^{\beta}$ and $\boldsymbol{Q}_{t+1,j}^{0}$. Moreover, since further covariance values are needed for $\mathcal{Q}_A$ and q_Q, it is necessary to compute all covariance values corresponding to the non-zeroes in $\left\langle \boldsymbol{Q}_{t,t+1}^{x} \right\rangle_{\setminus \mathcal{Q}_x}$. This can be achieved by applying sparse (reordered) Cholesky factorisation and partial matrix inversion that makes use of the Cholesky factor to solve the Takahashi equations. As expected all collapse operations on $q_{t,t+1}^{x}$ are connected, for this reason, we will detail each of them and then we assess how can we perform them in the most efficient way.

The Collapse$[q_{t,t+1}^{x}(\boldsymbol{x}_{t+1}); \boldsymbol{f}]$ *operation.* This operation involves projecting the the marginal density $q_{t,t+1}^{x}$ into the (sparse) Gaussian family defined by $\boldsymbol{f}$. The optimisation corresponding to this collapse step can be written as

(46)

Sparse approximations with known structure for $\boldsymbol{A}$ and $\boldsymbol{Q}$

In this section we discuss the sparsity issue that are essential for the scalability of the procedure we present. The basic assumptions we started with are that both $\boldsymbol{A}$ and $\boldsymbol{Q}$ are sparse. We also assume that $\boldsymbol{B}$ is a full matrix, although if we want to assume some structural sparsity, the issue can be addressed in the same way as for $\boldsymbol{A}$.

An crucial observation is that in all collapse steps we perform, we act on a sparse precision matrix and we compute all correlation values corresponding to the non-zero positions in this precision matrix. We argue that these correlation values are sufficient to keep the algorithm running, that is, there is no need to compute extra correlations and the correlations we compute always lead to positive semi-definite precision matrices. Given our observation about the computation of the correlation values, the first argument is quite intuitive: wherever there is a term with quadratic interaction between two variables of the same type be it $\boldsymbol{X}$, $\boldsymbol{A}$, $\boldsymbol{B}$ the moments of the other variables be it quadratic or just means are computed by their corresponding inference scheme and thus this property is conserved. The second argument related to positive (semi)-definiteness requires a closer look.

Firs we explore the connection between $\boldsymbol{X}$ and $\boldsymbol{A}$. We have $\left\langle \boldsymbol{Q}_{t,t+1}^{x} \right\rangle_{\setminus \mathcal{Q}_x} = \left\langle \boldsymbol{Q}_{t,t+1}^{x} \right\rangle_{q_0^A q_Q}$. Since $(\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t)^T \boldsymbol{Q} (\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t) \geq 0$ for any $\boldsymbol{x}_t, \boldsymbol{x}_{t+1}, \boldsymbol{A}$ and positive semi-definite $\boldsymbol{Q}$, it follows that $\left\langle (\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t)^T \left\langle \boldsymbol{Q} \right\rangle_{q_Q} (\boldsymbol{x}_{t+1} - \boldsymbol{A}\boldsymbol{x}_t) \right\rangle_{q_0^A} \geq 0$ for any positive semi-definite $\left\langle Q \right\rangle_{q_Q}$ thus $\left\langle \boldsymbol{Q}_{t,t+1}^{x} \right\rangle_{\setminus \mathcal{Q}_x}$ is positive semi-definite for all semi-definite $\left\langle Q \right\rangle_{q_Q}$. The argument for $\left\langle \boldsymbol{Q}_A \right\rangle_{\setminus \mathcal{Q}_A}$ is as follows: first we construct the matrix $\sum_t \left\langle \boldsymbol{x}_t \boldsymbol{x}_t^T \right\rangle_{\mathcal{Q}_x} \otimes \left\langle \boldsymbol{Q} \right\rangle_{q_Q} = \sum_t \left\langle \boldsymbol{x}_t \boldsymbol{x}_t^T \right\rangle_{q_{t,t+1}^x} \otimes \left\langle \boldsymbol{Q} \right\rangle_{q_Q}$ which, since it is a Kronecker product of positive semi-definite matrices, it is also positive semi-definite. Now, by eliminating the the rows and columns corresponding to the a-priori set structural zeros in $[\boldsymbol{A}]_c$ we arrive to a principal minor that also has to be positive semi-definite. The (i,j) block of $\left\langle \boldsymbol{Q}_A \right\rangle_{\setminus \mathcal{Q}_A}$ is $\left\langle x_t^i x_t^j \right\rangle [\left\langle Q \right\rangle_{q_Q}]_{I_i, I_j}$ where I_i and I_j are the supports of $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ respectively. Since the elements of $\boldsymbol{A}^T \boldsymbol{Q} \boldsymbol{A}$ are $\boldsymbol{a}_{I_i,i}^T \boldsymbol{Q}_{I_i, I_j} \boldsymbol{a}_{I_j, j}$, it follows that $\left\langle x_t^i x_t^j \right\rangle$ is always computed when needed, that is, when $I_i \cap I_j \neq \emptyset$. This shows that the sparsity structures can be exploited and that that the resulting inference keeps the matrices positive semi-definite.

A Appendix

A.1 Structured mean field approximation

$$\mathrm{D}[q_1(\boldsymbol{x}_{I_1})q_2(\boldsymbol{x}_{I_2})\dots q(\boldsymbol{x}_{I_K})||p(\boldsymbol{x})] = -\left\langle \log p(\boldsymbol{x})\right\rangle_{q_1 q_2 \dots q_K} + \sum_j \left\langle \log q_j(x_{I_j})\right\rangle q_j \tag{47}$$

$$q_j(\boldsymbol{x}_{I_j}) \propto \exp\{\langle \log p(\boldsymbol{x})\rangle_{\setminus q_j}\} \tag{48}$$

A.2 Inference using expectation constraints

Let us assume that we have a density that factors as

$$p(\boldsymbol{x}) = \frac{1}{Z_p}\prod_j \Psi_j(\boldsymbol{x}_{I_j})$$

and let us assume that the factors Ψ_j represent this density at the largest resolution, that is, no Ψ_j can be factored further. Now, say, we want to approximate p with a density q that has some desirable features, say, easy access to marginals or expectations (features common in exponential families), that p has not. The Kullback-Leibler divergence $\mathrm{D}[\cdot||\cdot]$ or some of its versions like the α-divergence come as natural measures to minimise. The $\mathrm{D}[p||q]$ version is typically out of question because it typically requires computations of expectations and marginals that we want to get around of in the first place. A few examples are: (1) say, we want to do a mean field approximation in the $\mathrm{D}\left[p(\boldsymbol{x})||\prod_j q(x_i)\right]$ sense, then, the optimality conditions lead to $q(x_i) = p(x_i)$ or (2) we want to q to have an exponential form $q(\boldsymbol{x};\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}\cdot\boldsymbol{f}(\boldsymbol{x}) - \log Z(\boldsymbol{\theta}))$, then we hve to compute $\langle \boldsymbol{f}_x\rangle_p$ followed by the moment transformation specific to this family. All these alternatives, involve the very steps we want to avoid.

However, the other version $\mathrm{D}[q||p]$ requires the computation of the entropy of q which can become complicated. Nonetheless, as a trade of we get a relatively easy access to p and we can exploit its factorisation. The latter, so called variational version of $\mathrm{D}[\cdot||\cdot]$ reads as

$$\mathrm{D}[q||p] = \langle \log q(\boldsymbol{x})\rangle_{q(\boldsymbol{x})} - \sum_j \left\langle \log \Psi_j(\boldsymbol{x}_{I_j})\right\rangle_{q(\boldsymbol{x}_{I_j})} + \log Z_p \tag{49}$$

In the above mentioned example we obtain the objectives

$$\mathrm{D}[q||p] = \sum_j \langle \log q_j(x_j)\rangle_{q_j(x)_j} - \sum_j \left\langle \log \Psi_j(\boldsymbol{x}_{I_j})\right\rangle_{\prod_{k\in I_j} q_k(x_k)} \tag{50}$$

and

$$\mathrm{D}[q||p] = \boldsymbol{\theta}\cdot\partial_{\boldsymbol{\theta}}\log Z(\boldsymbol{\theta}) - \log Z(\boldsymbol{\theta}) - \sum_j \left\langle \log \Psi_j(\boldsymbol{x}_{I_j})\right\rangle_{q(\boldsymbol{x}_{I_j})} \tag{51}$$

which are more manageable than the former version $\mathrm{D}[p||q]$.

In the following we'll try to go beyond the mean field approach in Equation (50). Let us assume that we have constructed an junction tree on the graph defined by the interactions in the cliques I_j, and let us denote the node sets of this tree by $C_1,\dots,C_K$ and let us use the notation $S_1,\dots,S_L$ for the separator corresponding to this tree. Moreover, let us restructure the factors $\Psi_j(\boldsymbol{x}_{I_j})$ by forming new factors $\Phi_j(\boldsymbol{x}_{C_j})$ corresponding to the clique structure of the junction tree. A factor $\Psi_j(\boldsymbol{x}_{I_j})$ can be attached to any factor $\Phi_j(\boldsymbol{x}_{C_j})$ where $I_k \subseteq C_j$ holds. We will assume that the approximating density follows the structure of the junction tree and thus can be written as

$$q(\boldsymbol{x}) = \frac{\prod_j q(\boldsymbol{x}_{C_j})}{\prod_k q(\boldsymbol{x}_{S_k})^{n_k-1}}.$$

where n_k is the number of cliques a separator appears in. Note that since S_k is separator $n_k \geq 2$. Since q has a clique-tree structure we can write the corresponding variational KL-divergence as

$$\mathrm{D}[q||p] = \sum_j \left\langle \log q(\boldsymbol{x}_{C_j}) \right\rangle_{q(\boldsymbol{x}_{C_j})} - \sum_k (n_k - 1) \left\langle \log q(\boldsymbol{x}_{S_k}) \right\rangle_{q(\boldsymbol{x}_{S_k})} - \sum_j \left\langle \log \Phi_j(\boldsymbol{x}_{C_j}) \right\rangle_{q(\boldsymbol{x}_{C_j})} + \log Z_p.$$

Clearly, in order to turn this into an optimisation problem we have to replace q by a family of approximate marginals $\mathcal{Q} = \{\{q_{C_j}(\boldsymbol{x}_{C_j})\}_j, \{q_{S_k}(\boldsymbol{x}_{S_k})\}_k\}$ where we assume that the constrains $\int d\boldsymbol{x}_{C_j \setminus S_k} q_{C_j}(\boldsymbol{x}_{C_j}) = q_{S_k}(\boldsymbol{x}_{S_k})$ hold for any $S_k \subseteq C_j$. With these constraints we define the variational objective

$$F(\mathcal{Q}) \equiv \sum_j \left\langle \log q_{C_j}(\boldsymbol{x}_{C_j}) \right\rangle_{q_{C_j}(\boldsymbol{x}_{C_j})} - \sum_k (n_k - 1) \left\langle \log q_{S_k}(\boldsymbol{x}_{S_k}) \right\rangle_{q_{S_k}(\boldsymbol{x}_{S_k})} - \sum_j \left\langle \log \Phi_j(\boldsymbol{x}_{C_j}) \right\rangle_{q_{C_j}(\boldsymbol{x}_{C_j})} \tag{52}$$

and the Lagrangian

$$\begin{aligned} L(\mathcal{Q}, \Lambda) \equiv \quad & F(\mathcal{Q}) + \sum_{j,k: S_k \subseteq C_j} \int d\boldsymbol{x}_{S_k}\, \mu_{j,k}(\boldsymbol{x}_{S_k}) [q_{S_k}(\boldsymbol{x}_{S_k}) - q_{C_j}(\boldsymbol{x}_{S_k})] \\ & + \sum_j z_{C_j} [\int d\boldsymbol{x}_{C_j}\, q_{C_j}(\boldsymbol{x}_{C_j}) - 1] + \sum_k z_{S_k} [1 - \int d\boldsymbol{x}_{S_k}\, q_{S_k}(\boldsymbol{x}_{S_k})]. \end{aligned}$$

From the stationarity conditions we obtain that

$$\begin{aligned} q_{C_j}(\boldsymbol{x}_{C_j}) &= \Phi_j(\boldsymbol{x}_{C_j}) \times \exp\Big\{ \sum_{k: S_k \subseteq C_j} \mu_{j,k}(\boldsymbol{x}_{S_k}) - z_{C_j} \Big\} \\ q_{S_k}(\boldsymbol{x}_{S_k}) &= \exp\Big\{ \frac{1}{n_k - 1} \sum_{j: C_j \supseteq S_k} \mu_{j,k}(\boldsymbol{x}_{S_k}) - z_{S_k}/(n_k - 1) \Big\} \\ q_{S_k}(\boldsymbol{x}_{S_k}) &= q_{C_j}(\boldsymbol{x}_{S_k}) \quad \text{for any} \quad C_j \cap S_k \neq \emptyset. \end{aligned}$$

The latter equation can be written as

$$\exp\Big\{ \frac{1}{n_k - 1} \sum_{j': C_j \supseteq S_k} \mu_{j',k}(\boldsymbol{x}_{S_k}) - z_{S_k}/(n_k - 1) \Big\} = \int d\boldsymbol{x}_{C_j \setminus S_k} \Phi_j(\boldsymbol{x}_{C_j}) \times \exp\Big\{ \sum_{k': S_k \subseteq C_j} \mu_{j,k'}(\boldsymbol{x}_{S_{k'}}) - z_{C_j} \Big\} \tag{53}$$

We introduce a non-singular linear transformation of the Lagrange multipliers

$$\mu_{jk}(\boldsymbol{x}_{S_k}) = \sum_{j' \neq j: C_{j'} \supseteq S_k} \hat{\mu}_{j'k}(\boldsymbol{x}_{S_k}).$$

By rearranging the terms we get

$$\begin{aligned} \exp\Big\{ \mu_{jk}(\boldsymbol{x}_{S_k}) + \hat{\mu}_{jk}(\boldsymbol{x}_{S_k}) \Big\} &\propto \int d\boldsymbol{x}_{C_j \setminus S_k} \Phi_j(\boldsymbol{x}_{C_j}) \times \exp\Big\{ \sum_{k': S_{k'} \subseteq C_j} \mu_{jk'}(\boldsymbol{x}_{S_{k'}}) \Big\} \\ &\propto \int d\boldsymbol{x}_{C_j \setminus S_k} \Phi_j(\boldsymbol{x}_{C_j}) \times \exp\Big\{ \sum_{k': S_{k'} \subseteq C_j} \sum_{j' \neq j: C_{j'} \supseteq S_k} \hat{\mu}_{j'k'}(\boldsymbol{x}_{S_{k'}}) \Big\}. \end{aligned}$$

Using these notations and stationarity conditions one can derive the fixed point equations/updates known as message passing. This has the form

$$q_{C_j}(\boldsymbol{x}_{C_j}) \propto \Phi_j(\boldsymbol{x}_{C_j}) \times \exp\Big\{ \sum_{k': S_{k'} \subseteq C_j} \mu_{jk'}(\boldsymbol{x}_{S_{k'}}) \Big\}, \tag{54}$$

$$q_{S_k}(\boldsymbol{x}_{S_k}) \propto \exp\Big\{ \sum_{j': C_{j'} \supseteq S_j} \hat{\mu}_{j'k}(\boldsymbol{x}_{S_k}) \Big\}, \tag{55}$$

$$\hat{\mu}_{jk}(\boldsymbol{x}_{S_k}) \doteq \log\Big\{ \int \boldsymbol{x}_{C_j \setminus S_k} q_{C_j}(\boldsymbol{x}_{C_j}) \Big\} - \mu_{jk}(\boldsymbol{x}_{S_k}), \tag{56}$$

$$\mu_{jk}(\boldsymbol{x}_{S_k}) \doteq \log\big\{ q_{S_k}(\boldsymbol{x}_{S_k}) \big\} - \hat{\mu}_{jk}(\boldsymbol{x}_{S_k}). \tag{57}$$

where we use the symbol "$\overset{\circ}{=}$" to denote equality up to a constant, for example, $2x \overset{\circ}{=} 2x + 1$. The functions μ_{jk} and $\hat{\mu}_{jk}$ are called messages. There are several specialisations of this algorithm, in the following we detail two of these: (1) when the constraints $q_{S_k}(\boldsymbol{x}_{S_k}) = q_{C_j}(\boldsymbol{x}_{S_k})$ is replaced by moment constraints and (2) when we no longer operate on clique-trees that is C_j and S_j are arbitrary.

A.2.1 Expectation constraints

In many cases the integrals in equations (53) and (61) are not analytically computable and having free form messages as shown above can be computationally demanding. To get a handle on this issue a plausible approach is to parameterise the messages (Lagrange multipliers). However, in many cases the marginal consistency constraints in (53) cannot be satisfied and further approximations have to be applied. First we show that parameterised Lagrange multipliers or messages are formally equivalent to moment matching constraints (instead of marginal matching constraints). Say, we want all messages to be in the linear parametric family defined by $\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{z})$. Then the Lagrange multiplier terms have the form

$$\int d\boldsymbol{x}_{S_k}[\boldsymbol{\mu}_{jk} \cdot \boldsymbol{f}(\boldsymbol{x}_{S_k})][q_{S_k}(\boldsymbol{x}_{S_k}) - q_{C_j}(\boldsymbol{x}_{S_k})] = \boldsymbol{\mu}_{jk} \cdot \big[\int d\boldsymbol{x}_{S_k} q_{S_k}(\boldsymbol{x}_{S_k}) \boldsymbol{f}(\boldsymbol{x}_{S_k}) - \int d\boldsymbol{x}_{S_k} q_{C_j}(\boldsymbol{x}_{S_k}) \boldsymbol{f}(\boldsymbol{x}_{S_k}) \big] \tag{58}$$

an thus this can be formally interpreted as replacing the marginal matching constraints $q_{C_j}(\boldsymbol{x}_{S_k}) = q_{S_k}(\boldsymbol{x}_{S_k})$ with the moment matching constraints $\langle f(\boldsymbol{x}_{S_k}) \rangle_{q_{C_j}} = \langle f(\boldsymbol{x}_{S_k}) \rangle_{q_{S_k}}$. This implies that the the marginal matching in Equation (53) is replaced my moment matching. Let us define

$$\text{Collapse}[p(\boldsymbol{z}); \boldsymbol{f}] = \operatorname*{argmin}_{\boldsymbol{\theta}} \text{D}[p(\boldsymbol{z}) || \exp(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{z}) - \log Z(\boldsymbol{\theta}))]$$

and assume that it is unique. Then the message passing algorithm can be rewritten as

$$q_{C_j}(\boldsymbol{x}_{C_j}) \propto \Phi_j(\boldsymbol{x}_{C_j}) \times \exp\Big\{ \sum_{k' : S_{k'} \subseteq C_j} \boldsymbol{\mu}_{jk'} \cdot \boldsymbol{f}(\boldsymbol{x}_{S_{k'}}) \Big\}, \tag{59}$$

$$q_{S_k}(\boldsymbol{x}_{S_k}) \propto \exp\Big\{ \sum_{j' : C_{j'} \supset S_j} \hat{\boldsymbol{\mu}}_{j'k} \cdot \boldsymbol{f}(\boldsymbol{x}_{S_k}) \Big\}, \tag{60}$$

$$\hat{\boldsymbol{\mu}}_{jk} = \text{Collapse}[q_{C_j}(\boldsymbol{x}_{S_k}); \boldsymbol{f}] - \boldsymbol{\mu}_{jk}, \tag{61}$$

$$\boldsymbol{\mu}_{jk} = \text{Collapse}[q_{S_k}(\boldsymbol{x}_{S_k}); \boldsymbol{f}] - \hat{\boldsymbol{\mu}}_{jk}. \tag{62}$$

Note that $q_{S_k}(\boldsymbol{x}_{S_k})$ is already in the family defined by $\boldsymbol{f}$, therefore, the only role of $\text{Collapse}[q_{S_k}(\boldsymbol{x}_{S_k}); \boldsymbol{f}]$ is parameter identification. It is important to mention that at no point the messages $\boldsymbol{\mu}_{jk}$ and $\hat{\boldsymbol{\mu}}_{jk}$ have to represent normalizable functions, that is, valid densities. It is sufficient to have q_{S_k} and q_{C_j} normalizable. Moreover, $\boldsymbol{f}$ can vary from constraint to constraint as long as manage to keep the computations under control.

A.2.2 Expectation propagation

A further algorithm follows by removing the clique-tree requirement, in this case we consider $C_j \cap C_j \not\subseteq C_j$ for any i and j and define the sets S_k as atomic sets satisfying the condition $S_k \cap C_j$ implies $S_k \subseteq C_j$. In this case we approximate the entropy by

$$\tilde{H}(\mathcal{Q}) = \sum_j \left\langle \log q(\boldsymbol{x}_{C_j}) \right\rangle_{q(\boldsymbol{x}_{C_j})} - \sum_k (n_k - 1) \left\langle \log q(\boldsymbol{x}_{S_k}) \right\rangle_{q(\boldsymbol{x}_{S_k})},$$

where n_k is again the number of sets of type C_j that S_k appears in, that is, $n_k = |\{C_j : S_k \subseteq S_j\}|$. This approximation of the entropy is similar to the Bethe entropy approximation in (?). The resulting algorithm is known as expectation propagation (?).