Derivations for NLARX node

Wouter. M Kouw

October 12, 2020

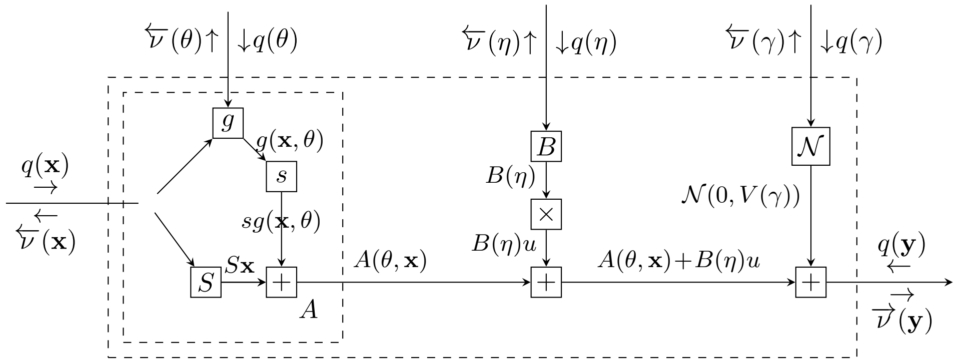

Figure 1: Nonlinear Latent AutoRegressive with eXogenous input node.

1 NLARX Node

The NLARX node is a Nonlinear Latent AutoRegressive model with eXogenous input. It is an autoregression on the latent states:

$$\mathbf{y} = \begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-M-1} \end{bmatrix} , \quad \mathbf{x} = \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-M} \end{bmatrix} . \tag{1}$$

The first element of $\mathbf{y}$ is predicted using a nonlinear stochastic autoregression on $\mathbf{x}$: $x_t = g(\mathbf{x}, \theta) + w_t$ where $w_t \sim \mathcal{N}(0, \gamma)$, θ are autoregression coefficients and g is a given nonlinear function. The remaining elements of $\mathbf{y}$ are a shifted down version of $\mathbf{x}$:

$$\begin{bmatrix} 0 \\ x_{t-1} \\ \vdots \\ x_{t-M-1} \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & \dots & \dots & 0 \\ 1 & 0 & \dots & \vdots \\ \vdots & \ddots & 0 & \vdots \\ 0 & \dots & 1 & 0 \end{bmatrix}}_{S} \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ \vdots \\ x_{t-M} \end{bmatrix} \tag{2}$$

Shifting is done to ensure that the model can be applied recursively over time. The node function has the following form:

$$\mathbf{y} \sim \mathcal{N}\big(A(\theta, \mathbf{x}) + B(\eta)u, V(\gamma)\big) \tag{3}$$

where $A(\theta, \mathbf{x}) = S\mathbf{x} + sg(\mathbf{x}, \theta)'$ with $s = \begin{bmatrix}1 & \dots & 0\end{bmatrix}^\top$. The term connected to the input variable u is $B(\eta) = s\eta$. $V(\gamma)$ is a matrix constructed from γ and a noise injection ϵ, which produces the following covariance and precision matrices:

$$V(\gamma) = \begin{bmatrix} \gamma^{-1} & 0 \\ 0 & \epsilon \end{bmatrix}, \qquad V^{-1}(\gamma) = W = \begin{bmatrix} \gamma & 0 \\ 0 & 1/\epsilon \end{bmatrix}. \tag{4}$$

The precision parameter γ is there for the stochastic autoregression. The (2,2) element of V should actually be 0 since the data shifting is deterministic. But a 0 element would cause the covariance matrix to be non-invertible. A common fix is to inject a small amount of noise ϵ (see [1]).

The derivations below are written for an autoregression of order 2, but it is straightforward to generalise this to higher orders (see also [2]).

2 Incoming messages

The messages that arrive at the node consist of the current recognition distributions for the unknown variables. I'm using a mean-field factorisation for the recognition model:

$$q(\mathbf{y}) = \mathcal{N}(\mathbf{y} \mid m_y, V_y) \tag{5a}$$
$$q(\theta) = \mathcal{N}(\theta \mid m_\theta, V_\theta) \tag{5b}$$
$$q(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid m_x, V_x) \tag{5c}$$
$$q(\eta) = \mathcal{N}(\eta \mid m_\eta, V_\eta) \tag{5d}$$
$$q(\gamma) = \Gamma(a_\gamma, b_\gamma)\,. \tag{5e}$$

Note that the expected value of the precision matrix W w.r.t. $q(\gamma)$ is:

$$\mathbb{E}_{q(\gamma)} W = m_W = \begin{bmatrix} m_\gamma & 0 \\ 0 & 1/\epsilon \end{bmatrix}, \tag{6}$$

with $m_\gamma = a_\gamma /\ b_\gamma$.

3 Outgoing messages

The general formula for computing an outgoing message is:

$$\nu(\mathbf{y}) = \exp\Big(\iint q(\mathbf{x})q(\theta) \log f(\mathbf{y}, \mathbf{x}, \theta) \ \ \mathrm{d}\mathbf{x}\mathrm{d}\theta \ \Big) \, . \tag{7}$$

In short: you take the expectation of the node function with respect to all unknown variables, except the one you're computing the message for. Below, I work out this formula for all outgoing messages of the NLARX node.

The following results are important in all message computations (for expectations with respect to the nonlinearity $g(\mathbf{x}, \theta)$, see Appendix 5):

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{x})q(\theta)} A(\theta, \mathbf{x}) &= \mathbb{E}_{q(\mathbf{x})q(\theta)} \big[S\mathbf{x} + sg(\mathbf{x}, \theta) \big] && \text{(8a)} \\
&= Sm_x + sg(m_x, m_\theta) && \text{(8b)} \\
&= A(m_\theta, m_x) \, . && \text{(8c)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{x})} A(\theta, \mathbf{x}) &= \mathbb{E}_{q(\mathbf{x})} \big[S\mathbf{x} + sg(\mathbf{x}, \theta) \big] && \text{(9a)} \\
&= Sm_x + s\big(g(m_x, m_\theta) + J_\theta^\top (\theta - m_\theta)\big) && \text{(9b)} \\
&= A(m_\theta, m_x) + s \ J_\theta^\top (\theta - m_\theta) \, . && \text{(9c)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{q(\theta)} A(\theta, \mathbf{x}) &= \mathbb{E}_{q(\theta)} \big[S\mathbf{x} + sg(\mathbf{x}, \theta) \big] && \text{(10a)} \\
&= S\mathbf{x} + sg(m_x, m_\theta) + sJ_\mathbf{x}^\top (\mathbf{x} - m_x) \, . && \text{(10b)}
\end{aligned}$$

3.1 Message to y

This is the outgoing message towards the next state.

$$
\begin{aligned}
&\log \overrightarrow{\nu}(\mathbf{y}) \\
&\quad= \iint q(\mathbf{x})q(\theta)q(\eta)q(\gamma) \log \mathcal{N}\big(A(\theta,\mathbf{x}) + B(\eta)u, V(\gamma)\big)\ \mathrm{d}\mathbf{x}\ \mathrm{d}\theta\ \mathrm{d}\eta\ \mathrm{d}\gamma && \text{(11a)}\\
&\quad\propto\ -\frac{1}{2}\ \mathbb{E}_q\big[(\mathbf{y} - A(\theta,\mathbf{x}) - B(\eta)u)^\top W(\mathbf{y} - A(\theta,\mathbf{x}) - B(\eta)u)\big] && \text{(11b)}\\
&\quad\propto\ -\frac{1}{2}\ \mathbb{E}_q\big[\mathbf{y}^\top W\mathbf{y} - \mathbf{y}^\top WA(\theta,\mathbf{x}) - \mathbf{y}^\top WB(\eta)u \\
&\qquad\qquad - A(\theta,\mathbf{x})^\top W\mathbf{y} - uB(\eta)^\top W\mathbf{y}\big] && \text{(11c)}\\
&\quad\propto\ -\frac{1}{2}\ \big[\mathbf{y}^\top m_W\mathbf{y} - \mathbf{y}^\top m_W A(m_\theta, m_{\mathbf{x}}) - \mathbf{y}^\top m_W B(m_\eta)u \\
&\qquad\qquad - A(m_\theta, m_{\mathbf{x}})^\top m_W\mathbf{y} - uB(m_\eta)^\top m_W\mathbf{y}\big] && \text{(11d)}\\
&\quad\propto\ -\frac{1}{2}\ \big[\mathbf{y}^\top \underbrace{m_W}_{\Phi}\mathbf{y} - \mathbf{y}^\top \underbrace{m_W\big(A(m_\theta, m_{\mathbf{x}}) + B(m_\eta)u\big)}_{\phi} \\
&\qquad\qquad - \underbrace{\big(A(m_\theta, m_{\mathbf{x}})^\top + uB(m_\eta)^\top\big)m_W}_{\phi^\top}\mathbf{y}\big] && \text{(11e)}\\
&\quad\propto \mathcal{N}(\mathbf{y} \mid \Phi^{-1}\phi,\ \Phi^{-1})\,. && \text{(11f)}
\end{aligned}
$$

3.2 Message to x

$$
\begin{aligned}
&\log \overleftarrow{\nu}(\mathbf{x}) \\
&\quad= \iint q(\mathbf{y})q(\theta)q(\eta)q(\gamma) \log \mathcal{N}\big(A(\theta,\mathbf{x}) + B(\eta)u, V(\gamma)\big)\ \mathrm{d}\mathbf{y}\ \mathrm{d}\theta\ \mathrm{d}\eta\ \mathrm{d}\gamma && \text{(12)}\\
&\quad\propto\ -\frac{1}{2}\ \mathbb{E}_q\big[(\mathbf{y} - A(\theta,\mathbf{x}) - B(\eta)u)^\top W(\mathbf{y} - A(\theta,\mathbf{x}) - B(\eta)u)\big] && \text{(13)}\\
&\quad\propto\ -\frac{1}{2}\ \mathbb{E}_q\big[- \mathbf{y}^\top WA(\theta,\mathbf{x}) + A(\theta,\mathbf{x})^\top WA(\theta,\mathbf{x}) \\
&\qquad\qquad + uB(\eta)^\top WA(\theta,\mathbf{x}) - A(\theta,\mathbf{x})^\top W\mathbf{y} + A(\theta,\mathbf{x})^\top WB(\eta)u\big] && \text{(14)}\\
&\quad\propto\ -\frac{1}{2}\ \big[- m_y^\top m_W(Sx + sg(m_x, m_\theta) + sJ_{\mathbf{x}}^\top(\mathbf{x} - m_x)) \\
&\qquad\qquad + \underbrace{\mathbb{E}_{q(\theta)}\big[A(\theta,\mathbf{x})^\top m_W A(\theta,\mathbf{x})\big]}_{\text{(I)}} \\
&\qquad\qquad + um_\eta s^\top m_W(S\mathbf{x} + sg(m_x, m_\theta) + sJ_{\mathbf{x}}^\top(\mathbf{x} - m_x)) \\
&\qquad\qquad - (S\mathbf{x} + sg(m_x, m_\theta) + sJ_{\mathbf{x}}^\top(\mathbf{x} - m_x))^\top m_W m_y \\
&\qquad\qquad + (S\mathbf{x} + sg(m_x, m_\theta) + sJ_{\mathbf{x}}^\top(\mathbf{x} - m_x))^\top m_W s m_\eta u\big]\,. && \text{(15)}
\end{aligned}
$$

Separate derivation for (I):

$$
\begin{aligned}
\text{(I)} &= \mathbb{E}_{q(\theta)}\big[A(\theta,\mathbf{x})^\top m_W A(\theta,\mathbf{x})\big] \\
&= \mathbb{E}_{q(\theta)}\big[(S\mathbf{x}+sg(\mathbf{x},\theta))^\top m_W (S\mathbf{x}+sg(\mathbf{x},\theta))\big] &\text{(16a)}\\
&= \mathbb{E}_{q(\theta)}\big[(S\mathbf{x})^\top m_W S\mathbf{x} + (S\mathbf{x})^\top m_W sg(\mathbf{x},\theta) \\
&\qquad + (sg(\mathbf{x},\theta))^\top m_W S\mathbf{x} + (sg(\mathbf{x},\theta))^\top m_W sg(\mathbf{x},\theta)\big] &\text{(16b)}\\
&= \mathbf{x}^\top S^\top m_W S\mathbf{x} \\
&\qquad + \mathbf{x}^\top \underbrace{S^\top m_W s}_{=0}\big(g(m_x,m_\theta) + J_{\mathbf{x}}^\top(\mathbf{x}-m_x)\big) \\
&\qquad\quad + \big(g(m_x,m_\theta) + J_{\mathbf{x}}^\top(\mathbf{x}-m_x)\big)^\top \underbrace{s^\top m_W S}_{=0} x \\
&\qquad\qquad + \underbrace{\mathbb{E}_{q(\theta)}\big[g(\mathbf{x},\theta)^\top s^\top m_W sg(\mathbf{x},\theta)\big]}_{\text{(II)}}\,. &\text{(16c)}
\end{aligned}
$$

Separate derivation for (II):

$$
\begin{aligned}
\text{(II)} &= \mathbb{E}_{q(\theta)}\big[g(\mathbf{x},\theta)^\top \underbrace{s^\top m_W s}_{m_\gamma} g(\mathbf{x},\theta)\big] &\text{(17a)}\\
&= m_\gamma \mathbb{E}_{q(\theta)}\big[g(\mathbf{x},\theta)^\top g(\mathbf{x},\theta)\big] &\text{(17b)}\\
&\propto m_\gamma\big((g(m_x,m_\theta) - J_{\mathbf{x}}^\top m_x)J_{\mathbf{x}}^\top \mathbf{x} \\
&\qquad + \mathbf{x}^\top J_{\mathbf{x}}(g(m_x,m_\theta) - J_{\mathbf{x}}^\top m_x) + \mathbf{x}^\top J_{\mathbf{x}} J_{\mathbf{x}}^\top \mathbf{x}\big) &\text{(17c)}
\end{aligned}
$$

The expectation of $g(\mathbf{x},\theta)^2$ with respect to $q(\theta)$ is described in Section 5.2.3. Putting (II) back in (I):

$$
\begin{aligned}
\text{(I)} &\propto (S\mathbf{x})^\top m_W S\mathbf{x} + m_\gamma\big((g(m_x,m_\theta) - J_{\mathbf{x}}^\top m_x)J_{\mathbf{x}}^\top \mathbf{x} \\
&\qquad + \mathbf{x}^\top J_{\mathbf{x}}(g(m_x,m_\theta) - J_{\mathbf{x}}^\top m_x) + \mathbf{x}^\top J_{\mathbf{x}} J_{\mathbf{x}}^\top \mathbf{x}\big) &\text{(18a)}\\
&= \mathbf{x}^\top\big(S^\top m_W S + m_\gamma J_{\mathbf{x}} J_{\mathbf{x}}^\top\big)\mathbf{x} \\
&\qquad + m_\gamma\big((g(m_x,m_\theta) - J_{\mathbf{x}}^\top m_x)J_{\mathbf{x}}^\top \mathbf{x} \\
&\qquad\quad + \mathbf{x}^\top J_{\mathbf{x}}(g(m_x,m_\theta) - J_{\mathbf{x}}^\top m_x) m_\gamma\,. &\text{(18b)}
\end{aligned}
$$

Putting (I) back in:

$$
\begin{aligned}
&\log \overleftarrow{\nu}(\mathbf{x}) \\
&\propto\; -\frac{1}{2}\big[-(m_y - s m_\eta u)^\top m_W (S\mathbf{x} + s g(m_x, m_\theta) + s J_{\mathbf{x}}^\top (\mathbf{x} - m_x)) \\
&\qquad + \mathbf{x}^\top \big(S^\top m_W S + m_\gamma J_{\mathbf{x}} J_{\mathbf{x}}^\top\big)\mathbf{x} \\
&\qquad + m_\gamma\big((g(m_x, m_\theta) - J_{\mathbf{x}}^\top m_x) J_{\mathbf{x}}^\top \mathbf{x} \\
&\qquad + \mathbf{x}^\top J_{\mathbf{x}} (g(m_x, m_\theta) - J_{\mathbf{x}}^\top m_x) m_\gamma \\
&\qquad\quad - (S\mathbf{x} + s g(m_x, m_\theta) + s J_{\mathbf{x}}^\top (\mathbf{x} - m_x))^\top m_W (m_y - s m_\eta u)\big] && \text{(19a)} \\
&\propto\; -\frac{1}{2}\big[-(m_y - s m_\eta u)^\top m_W (S + s J_{\mathbf{x}}^\top)\mathbf{x} \\
&\qquad + \mathbf{x}^\top \big(S^\top m_W S + m_\gamma J_{\mathbf{x}} J_{\mathbf{x}}^\top\big)\mathbf{x} \\
&\qquad\quad + m_\gamma\big((g(m_x, m_\theta) - J_{\mathbf{x}}^\top m_x) J_{\mathbf{x}}^\top \mathbf{x} \\
&\qquad\qquad + \mathbf{x}^\top J_{\mathbf{x}} (g(m_x, m_\theta) - J_{\mathbf{x}}^\top m_x) m_\gamma \\
&\qquad\qquad\quad - \mathbf{x}^\top (S + s J_{\mathbf{x}}^\top)^\top m_W (m_y - s m_\eta u)\big] && \text{(19b)} \\
&\propto\; -\frac{1}{2}\big[-\Big((m_y - s m_\eta u)^\top m_W (S + s J_{\mathbf{x}}^\top) - m_\gamma\big((g(m_x, m_\theta) - J_{\mathbf{x}}^\top m_x) J_{\mathbf{x}}^\top\Big)\mathbf{x} \\
&\qquad - \mathbf{x}^\top \underbrace{\Big((S + s J_{\mathbf{x}}^\top)^\top m_W (m_y - s m_\eta u) - J_{\mathbf{x}} (g(m_x, m_\theta) - J_{\mathbf{x}}^\top m_x) m_\gamma\Big)}_{\phi} \\
&\qquad + \mathbf{x}^\top \underbrace{\big(S^\top m_W S + m_\gamma J_{\mathbf{x}} J_{\mathbf{x}}^\top\big)}_{\Phi}\mathbf{x}\big] && \text{(19c)} \\
&\propto\; \mathcal{N}(\mathbf{x} \mid \Phi^{-1}\phi,\; \Phi^{-1})\,. && \text{(19d)}
\end{aligned}
$$

3.3 Message to θ

$$
\begin{aligned}
&\log \overleftarrow{\nu}(\theta) \\
&\quad = \iint q(\mathbf{y})q(\mathbf{x})q(\eta)q(\gamma) \log \mathcal{N}\big(A(\theta,\mathbf{x}) + B(\eta)u, V(\gamma)\big)\ \mathrm{d}\mathbf{y}\ \mathrm{d}\mathbf{x}\ \mathrm{d}\eta\ \mathrm{d}\gamma \quad (20) \\
&\quad \propto -\frac{1}{2}\ \mathbb{E}_q\big[(\mathbf{y} - A(\theta,\mathbf{x}) - B(\eta)u)^\top W(\mathbf{y} - A(\theta,\mathbf{x}) - B(\eta)u)\big] \quad (21) \\
&\quad \propto -\frac{1}{2}\ \mathbb{E}_q\big[- \mathbf{y}^\top W A(\theta,\mathbf{x}) + A(\theta,\mathbf{x})^\top W A(\theta,\mathbf{x}) \\
&\qquad\qquad + uB(\eta)^\top W A(\theta,\mathbf{x}) - A(\theta,\mathbf{x})^\top W\mathbf{y} + A(\theta,\mathbf{x})^\top W B(\eta)u\big] \quad (22) \\
&\quad \propto -\frac{1}{2}\ \big[- m_y^\top m_W (Sm_x + sg(m_x,m_\theta) + sJ_\theta^\top(\theta - m_\theta)) \\
&\qquad\qquad + \underbrace{\mathbb{E}_{q(\mathbf{x})}\big[A(\theta,\mathbf{x})^\top m_W A(\theta,\mathbf{x})\big]}_{\text{(I)}} \\
&\qquad\qquad + um_\eta s^\top m_W (Sm_x + sg(m_x,m_\theta) + sJ_\theta^\top(\theta - m_\theta)) \\
&\qquad\qquad - (Sm_x + sg(m_x,m_\theta) + sJ_\theta^\top(\theta - m_\theta))^\top m_W m_y \\
&\qquad\qquad + (Sm_x + sg(m_x,m_\theta) + sJ_\theta^\top(\theta - m_\theta))^\top m_W s m_\eta u\big]\,. \quad (23)
\end{aligned}
$$

Separate derivation for (I):

$$
\begin{aligned}
\text{(I)} &= \mathbb{E}_{q(\mathbf{x})}\big[A(\theta,\mathbf{x})^\top m_W A(\theta,\mathbf{x})\big] \\
&= \mathbb{E}_{q(\mathbf{x})}\big[(S\mathbf{x} + sg(\mathbf{x},\theta))^\top m_W (S\mathbf{x} + sg(\mathbf{x},\theta))\big] \quad (24a) \\
&= \mathbb{E}_{q(\mathbf{x})}\big[(S\mathbf{x})^\top m_W S\mathbf{x} + (S\mathbf{x})^\top m_W sg(\mathbf{x},\theta) \\
&\qquad + (sg(\mathbf{x},\theta))^\top m_W S\mathbf{x} + (sg(\mathbf{x},\theta))^\top m_W sg(\mathbf{x},\theta)\big] \quad (24b) \\
&= \mathbb{E}_{q(\mathbf{x})}[\mathbf{x}^\top S^\top m_W S\mathbf{x}] \\
&\qquad + \mathbb{E}_{q(\mathbf{x})}[\mathbf{x}^\top \underbrace{S^\top m_W s}_{=0} \big(g(m_x,m_\theta) + J_\theta^\top(\theta - m_\theta)\big)] \\
&\qquad\qquad + \mathbb{E}_{q(\mathbf{x})}[\big(g(m_x,m_\theta) + J_\theta^\top(\theta - m_\theta)\big)^\top \underbrace{s^\top m_W S}_{=0}\mathbf{x}] \\
&\qquad\qquad\quad + \mathbb{E}_{q(\mathbf{x})}\big[g(\mathbf{x},\theta)^\top s^\top m_W sg(\mathbf{x},\theta)\big] \quad (24c) \\
&\propto \mathbb{E}_{q(\mathbf{x})}\big[g(\mathbf{x},\theta)^\top \underbrace{s^\top m_W s}_{m_\gamma} g(\mathbf{x},\theta)\big] \quad (24d) \\
&= m_\gamma \mathbb{E}_{q(\mathbf{x})}\big[g(\mathbf{x},\theta)^\top g(\mathbf{x},\theta)\big] \quad (24e) \\
&\propto m_\gamma\big((g(m_x,m_\theta) - J_\theta^\top m_\theta)J_\theta^\top\theta \\
&\qquad + \theta^\top J_\theta(g(m_x,m_\theta) - J_\theta^\top m_\theta) + \theta^\top J_\theta J_\theta^\top\theta\big)\,. \quad (24f)
\end{aligned}
$$

Putting (I) back into the message derivation:

$$
\begin{aligned}
&\log \overleftarrow{\nu}(\theta) \\
&\quad \propto\ -\frac{1}{2}\ \big[-(m_y - s m_\eta u)^\top m_W (S m_x + s g(m_x, m_\theta) + s J_\theta^\top (\theta - m_\theta)) \\
&\qquad\qquad + m_\gamma (g(m_x, m_\theta) - J_\theta^\top m_\theta) J_\theta^\top \theta \\
&\qquad\qquad + m_\gamma \theta^\top J_\theta (g(m_x, m_\theta) - J_\theta^\top m_\theta) \\
&\qquad\qquad + m_\gamma \theta^\top J_\theta J_\theta^\top \theta \\
&\qquad\qquad - (S m_x + s g(m_x, m_\theta) + s J_\theta^\top (\theta - m_\theta))^\top m_W (m_y - s m_\eta u) \big]
\end{aligned}
\tag{25a}
$$

$$
\begin{aligned}
&\quad \propto\ -\frac{1}{2}\ \big[-(m_y - s m_\eta u)^\top m_W s J_\theta^\top \theta \\
&\qquad\qquad + m_\gamma (g(m_x, m_\theta) - J_\theta^\top m_\theta) J_\theta^\top \theta \\
&\qquad\qquad + m_\gamma \theta^\top J_\theta (g(m_x, m_\theta) - J_\theta^\top m_\theta) \\
&\qquad\qquad + m_\gamma \theta^\top J_\theta J_\theta^\top \theta \\
&\qquad\qquad - (s J_\theta^\top \theta)^\top m_W (m_y - s m_\eta u) \big]
\end{aligned}
\tag{25b}
$$

$$
\begin{aligned}
&\quad \propto\ -\frac{1}{2}\ \big[-\big((m_y - s m_\eta u)^\top m_W s - m_\gamma (g(m_x, m_\theta) - J_\theta^\top m_\theta)\big) J_\theta^\top \theta \\
&\qquad\qquad - \theta^\top \underbrace{\big(J_\theta s^\top m_W (m_y - s m_\eta u) - m_\gamma J_\theta (g(m_x, m_\theta) - J_\theta^\top m_\theta)\big)}_{\phi} \\
&\qquad\qquad + \theta^\top \underbrace{m_\gamma J_\theta J_\theta^\top}_{\Phi} \theta \\
&\quad \propto \mathcal{N}(\theta \mid \Phi^{-1} \phi,\ \Phi^{-1}) .
\end{aligned}
\tag{25c}
$$

3.4 Message to η

$$
\begin{aligned}
&\log \overleftarrow{\nu}(\eta) \\
&\quad = \iint q(\mathbf{y})q(\mathbf{x})q(\theta)q(\gamma) \log \mathcal{N}\big(A(\theta,\mathbf{x}) + B(\eta)u, V(\gamma)\big) \ \mathrm{d}\mathbf{y}\ \mathrm{d}\mathbf{x}\ \mathrm{d}\theta\ \mathrm{d}\gamma && \text{(26a)} \\
&\quad \propto -\frac{1}{2}\ \mathbb{E}_q\big[(\mathbf{y} - A(\theta,\mathbf{x}) - B(\eta)u)^\top W(\mathbf{y} - A(\theta,\mathbf{x}) - B(\eta)u)\big] && \text{(26b)} \\
&\quad \propto -\frac{1}{2}\ \mathbb{E}_q\big[- \mathbf{y}^\top W B(\eta)u + A(\theta,\mathbf{x})^\top W B(\eta)u \\
&\qquad\qquad + uB(\eta)^\top W B(\eta)u - uB(\eta)^\top W\mathbf{y} + uB(\eta)^\top W A(\theta,\mathbf{x})\big] && \text{(26c)} \\
&\quad \propto -\frac{1}{2}\ \big[- m_y^\top m_W us\eta + A(m_\theta, m_\mathbf{x})^\top m_W us\eta \\
&\qquad\qquad + \eta us^\top m_W su\eta - \eta us^\top m_W m_y + u\eta s^\top m_W A(m_\theta, m_\mathbf{x})\big] && \text{(26d)} \\
&\quad \propto -\frac{1}{2}\ \big[- (m_y - A(m_\theta, m_\mathbf{x}))^\top m_W su\eta \\
&\qquad\qquad - \eta \underbrace{us^\top m_W (m_y - A(m_\theta, m_\mathbf{x}))}_{\phi}\ + \eta \underbrace{us^\top m_W su}_{\Phi} \eta\big] && \text{(26e)} \\
&\quad \propto\ \mathcal{N}(\eta \mid \Phi^{-1}\phi,\ \Phi^{-1})\,. && \text{(26f)}
\end{aligned}
$$

3.5 Message to γ

$$
\begin{aligned}
&\log \overleftarrow{\nu}(\gamma) \\
&\quad = \iint q(\mathbf{y})q(\mathbf{x})q(\theta)q(\eta) \log \mathcal{N}\big(A(\theta,\mathbf{x}) + B(\eta)u, V(\gamma)\big) \ \mathrm{d}\mathbf{y}\ \mathrm{d}\mathbf{x}\ \mathrm{d}\theta\ \mathrm{d}\eta && \text{(27)} \\
&\quad \propto \frac{1}{2}\log|W| \\
&\qquad - \frac{1}{2}\ \mathrm{tr}(W\mathbb{E}_q\big[(\mathbf{y} - A(\theta,\mathbf{x}) - B(\eta)u)(\mathbf{y} - A(\theta,\mathbf{x}) - B(\eta)u)^\top\big] && \text{(28)} \\
&\quad \propto \frac{1}{2}\log|W| \\
&\qquad - \frac{1}{2}\ \mathrm{tr}(W\mathbb{E}_q\big[\underbrace{\mathbf{y}\mathbf{y}^\top}_{(1)} - \underbrace{A(\theta,\mathbf{x})\mathbf{y}^\top}_{(2)} - \underbrace{B(\eta)u\mathbf{y}^\top}_{(3)} \\
&\qquad - \underbrace{\mathbf{y}A(\theta,\mathbf{x})^\top}_{(4)} + \underbrace{A(\theta,\mathbf{x})A(\theta,\mathbf{x})^\top}_{(5)} + \underbrace{B(\eta)uA(\theta,\mathbf{x})^\top}_{(6)} \\
&\qquad - \underbrace{\mathbf{y}(B(\eta)u)^\top}_{(7)} + \underbrace{A(\theta,\mathbf{x})(B(\eta)u)^\top}_{(8)} + \underbrace{B(\eta)u(B(\eta)u)^\top}_{(9)}\big]\,. && \text{(29)}
\end{aligned}
$$

Each of these terms will be derived separately:

$$
\begin{aligned}
(1) &= \mathbb{E}_{q(\mathbf{y})}\big[\mathbf{y}\mathbf{y}^\top\big] && \text{(30a)} \\
&= m_y m_y^\top + V_y\,, && \text{(30b)}
\end{aligned}
$$

$$
\begin{aligned}
(2) &= \mathbb{E}_{q(\mathbf{y})q(\theta)q(\mathbf{x})}\big[A(\theta,\mathbf{x})\mathbf{y}^\top\big] && \text{(31a)}\\
&= A(m_\theta, m_x)m_y^\top\,, && \text{(31b)}
\end{aligned}
$$

$$
\begin{aligned}
(3) &= \mathbb{E}_{q(\mathbf{y})q(\eta)}\big[B(\eta)u\mathbf{y}^\top\big] && \text{(32a)}\\
&= B(m_\eta)um_y^\top\,, && \text{(32b)}
\end{aligned}
$$

$$
\begin{aligned}
(4) &= \mathbb{E}_{q(\mathbf{y})q(\theta)q(\mathbf{x})}\big[\mathbf{y}A(\theta,\mathbf{x})^\top\big] && \text{(33a)}\\
&= m_yA(m_\theta, m_x)^\top\,, && \text{(33b)}
\end{aligned}
$$

$$
\begin{aligned}
(5) &= \mathbb{E}_{q(\theta)q(\mathbf{x})}\big[A(\theta,\mathbf{x})A(\theta,\mathbf{x})^\top\big] && \text{(34a)}\\
&= \mathbb{E}_{q(\theta)q(\mathbf{x})}\big[(S\mathbf{x}+sg(\mathbf{x},\theta))(S\mathbf{x}+sg(\mathbf{x},\theta)^\top\big] && \text{(34b)}\\
&= \mathbb{E}_{q(\theta)q(\mathbf{x})}\big[S\mathbf{x}\mathbf{x}^\top S^\top + S\mathbf{x}g(\mathbf{x},\theta)^\top s^\top \\
&\qquad + sg(\mathbf{x},\theta)\mathbf{x}^\top S^\top + sg(\mathbf{x},\theta)g(\mathbf{x},\theta)s^\top\big] && \text{(34c)}\\
&= S(m_xm_x^\top + V_x)S^\top \\
&\quad + S\,\underbrace{\mathbb{E}_{q(\mathbf{x})q(\theta)}[\mathbf{x}g(\mathbf{x},\theta)^\top]}_{(I)}\,s^\top \\
&\quad + s\mathbb{E}_{q(\mathbf{x})q(\theta)}[g(\mathbf{x},\theta)\mathbf{x}^\top]S^\top \\
&\quad + s(g(m_x,m_\theta)^2 + J_\mathbf{x}^\top V_\mathbf{x}J_\mathbf{x} + J_\theta^\top V_\theta J_\theta)s^\top && \text{(34d)}\\
(I) &= \mathbb{E}_{q(\mathbf{x})q(\theta)}[\mathbf{x}g(\mathbf{x},\theta)^\top] && \text{(34e)}\\
&= \mathbb{E}_{q(\mathbf{x})q(\theta)}[\mathbf{x}(g(m_\mathbf{x},m_\theta) + J_\mathbf{x}^\top(\mathbf{x}-m_x) + J_\theta^\top(\theta-m_\theta))] && \text{(34f)}\\
&= \mathbb{E}_{q(\mathbf{x})}[\mathbf{x}(g(m_\mathbf{x},m_\theta) + J_\mathbf{x}^\top(\mathbf{x}-m_x))] && \text{(34g)}\\
&= \mathbb{E}_{q(\mathbf{x})}[\mathbf{x}g(m_\mathbf{x},m_\theta) + \mathbf{x}J_\mathbf{x}^\top\mathbf{x} - \mathbf{x}J_\mathbf{x}^\top m_x] && \text{(34h)}\\
&= m_xg(m_x,m_\theta) + \mathbb{E}_{q(\mathbf{x})}[\mathbf{x}\mathbf{x}^\top]J_\mathbf{x} - m_xm_x^\top J_\mathbf{x} && \text{(34i)}\\
&= m_xg(m_\mathbf{x},m_\theta) + (m_xm_x^\top + V_\mathbf{x})J_\mathbf{x} - m_xm_x^\top J_\mathbf{x} && \text{(34j)}\\
&= m_xg(m_\mathbf{x},m_\theta) + V_xJ_\mathbf{x} && \text{(34k)}\\
\Longrightarrow\ (5) &= S(m_xm_x^\top + V_x)S^\top \\
&\quad + S(m_xg(m_x,m_\theta) + V_xJ_\mathbf{x})s^\top \\
&\quad + s(m_xg(m_x,m_\theta) + V_xJ_\mathbf{x})^\top S^\top \\
&\quad + s(g(m_x,m_\theta)^2 + J_\mathbf{x}^\top V_xJ_\mathbf{x} + J_\theta^\top V_\theta J_\theta)s^\top\,. && \text{(34l)}
\end{aligned}
$$

$$
\begin{aligned}
(6) &= \mathbb{E}_{q(\eta)q(\theta)q(\mathbf{x})}\big[B(\eta)uA(\theta,\mathbf{x})^\top\big] && \text{(35a)}\\
&= B(m_\eta)uA(m_\theta, m_x)^\top\,, && \text{(35b)}
\end{aligned}
$$

$$
\begin{aligned}
(7) &= \mathbb{E}_{q(\mathbf{y})q(\eta)}\big[\mathbf{y}(B(\eta)u)^\top\big] && \text{(36a)}\\
&= m_y(B(m_\eta)u)^\top, && \text{(36b)}
\end{aligned}
$$

$$
\begin{aligned}
(8) &= \mathbb{E}_{q(\eta)q(\theta)q(\mathbf{x})}\big[A(\theta,\mathbf{x})(B(\eta)u)^\top\big] && (37a)\\
&= A(m_\theta, m_x)(B(m_\eta)u)^\top\,, && (37b)
\end{aligned}
$$

$$
\begin{aligned}
(9) &= \mathbb{E}_{q(\eta)}\big[B(\eta)u(B(\eta)u)^\top\big] && (38a)\\
&= u^2\mathbb{E}_{q(\eta)}\big[s\eta\eta s^\top\big] && (38b)\\
&= u^2 s(m_\eta^2 + V_\eta)s^\top\,. && (38c)
\end{aligned}
$$

In total, the message towards γ is:

$$
\begin{aligned}
\log \overleftarrow{\nu}(\gamma) &\propto \frac{1}{2}\log|W| - \frac{1}{2}\mathrm{tr}(W\Phi) && (39a)\\
&\propto \frac{1}{2}\log\gamma - \frac{1}{2}\gamma\Phi^{(1,1)} && (39b)\\
&\propto \Gamma(\frac{3}{2}, \frac{1}{2}\Phi^{(1,1)})\,, && (39c)
\end{aligned}
$$

where $\Phi = (1)-(2)-(3)-(4)+(5)+(6)-(7)+(8)+(9)$. Note that the trace of a product of matrices AB is $A^{(1,1)}\cdot B^{(1,1)} + A^{(2,2)}\cdot B^{(2,2)} + \ldots$ and since γ only appears in the (1,1) element of W, the result of the trace $\mathrm{tr}(W\Phi)$ is proportional to $\gamma\Phi^{(1,1)}$. Same as true for the log-determinant of W: $\log W^{(1,1)} + \log W^{2,2} + \cdots \propto \log\gamma$.

4 Free energy

Each node has an "energy" term (from the energy-entropy decomposition of free energy), referred to as $U[q]$. Shifting the data is only necessary so the model can be applied recursively, but it is not part of the prediction for the current state x_t. To compute free energy, we must ignore the data shifting and just consider the regression of x_t from $x_{t-1}, \dots, x_{t-M}$. For the NLARX node, it is:

$$U[q] = \\ - \int q(\mathbf{y})q(\mathbf{x})q(\theta)q(\eta)q(\gamma) \log \mathcal{N}\big(\mathbf{y}^{(1)} \mid g(\mathbf{x}, \theta) + \eta u, \gamma\big) \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x}\mathrm{d}\theta\mathrm{d}\eta\mathrm{d}\gamma \tag{40}$$

where $\mathbf{y}^{(1)}$ is x_t. The free energy is:

$$\begin{aligned}
U[q] &= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\mathbb{E}_{q(\gamma)}[\log \gamma] - \frac{1}{2}\mathbb{E}_q\big[\gamma(\mathbf{y}^{(1)} - g(\mathbf{x},\theta) - \eta u)^2\big] && \text{(41a)} \\
&= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\big(\psi(a_\gamma)) - \log(b_\gamma)\big) \\
&\quad - \frac{1}{2}\frac{a_\gamma}{b_\gamma}\mathbb{E}_q\big[\mathbf{y}^{(1)}\mathbf{y}^{(1)} - \mathbf{y}^{(1)}g(\mathbf{x},\theta) - \mathbf{y}^{(1)}\eta u - g(\mathbf{x},\theta)\mathbf{y}^{(1)} \\
&\qquad + g(\mathbf{x},\theta)^2 + g(\mathbf{x},\theta)\eta u - \eta u\mathbf{y}^{(1)} + \eta u g(\mathbf{x},\theta)(\eta u)^2\big] && \text{(41b)} \\
&= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\big(\psi(a_\gamma)) - \log(b_\gamma)\big) \\
&\quad - \frac{1}{2}\frac{a_\gamma}{b_\gamma}\Big[m_y^{(1)}m_y^{(1)} + V_y^{(1,1)} - m_y^{(1)}g(m_x, m_\theta) \\
&\qquad - m_y^{(1)}m_\eta u - g(m_x, m_\theta)m_y^{(1)} + g(m_x, m_\theta)^2 \\
&\qquad + J_{\mathbf{x}}^\top V_x J_{\mathbf{x}} + J_\theta^\top V_\theta J_\theta + g(m_x, m_\theta)m_\eta u \\
&\qquad - m_\eta u m_y^{(1)} + m_\eta u g(m_x, m_\theta) + (m_\eta^2 + V_\eta)u^2\Big]. && \text{(41c)}
\end{aligned}$$

5 Appendix: Taylor approx. of nonlinearity

In the message computations listed below, I will make use of a Taylor approximation of the nonlinearity g. The approximation is first-order in both arguments and uses the means of the recognition distributions (m_x, m_θ) as approximating points:

$$g(\mathbf{x}, \theta) = g(m_x, m_\theta) + J_{\mathbf{x}}^\top (\mathbf{x} - m_x) + J_\theta^\top (\theta - m_\theta)\,, \tag{42}$$

where $J_{\mathbf{x}}$ denotes the partial derivative of g with respect to x, i.e.

$$J_{\mathbf{x}} \triangleq \left. \frac{\partial g(\mathbf{x}, \theta)}{\partial x} \right|_{\mathbf{x}=m_x, \theta=m_\theta}\,, \tag{43}$$

and likewise J_θ w.r.t. θ.

5.1 First-order moments

Throughout the message computations, we will need the expected values of g, possibly raised to some power, w.r.t. the recognition distributions $q(\mathbf{x})$ and $q(\theta)$. Note that in message computations we sometimes want to drop terms that do not depend on the particular variable of interest. For instance, in computing the message $\overleftarrow{\nu}(\theta)$, we need $\mathbb{E}_{q(\mathbf{x})}[g(m_x, m_\theta)]$ but can ignore all terms that don't involve θ. In that case, the final line of the derivations below will have a "proportional to" sign $\propto$, where terms have been dropped.

5.1.1 With respect to θ and x

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{x}),q(\theta)}[g(\mathbf{x}, \theta)] &= \mathbb{E}_{q(\mathbf{x}),q(\theta)}[g(m_x, m_\theta) + J_{\mathbf{x}}^\top (\mathbf{x} - m_x) + J_\theta^\top (\theta - m_\theta)] && \text{(44a)} \\
&= g(m_x, m_\theta) + J_{\mathbf{x}}^\top \underbrace{\mathbb{E}_{q(\mathbf{x})}[(\mathbf{x} - m_x)]}_{0} + J_\theta^\top \underbrace{\mathbb{E}_{q(\theta)}[(\theta - m_\theta)]}_{0} && \text{(44b)} \\
&= g(m_x, m_\theta)\,. && \text{(44c)}
\end{aligned}$$

5.1.2 With respect to x

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{x})}[g(\mathbf{x}, \theta)] &= \mathbb{E}_{q(\mathbf{x})}[g(m_x, m_\theta) + J_{\mathbf{x}}^\top (\mathbf{x} - m_x) + J_\theta^\top (\theta - m_\theta)] && \text{(45a)} \\
&= g(m_x, m_\theta) + J_{\mathbf{x}}^\top \underbrace{\mathbb{E}_{q(\mathbf{x})}[(\mathbf{x} - m_x)]}_{0} + J_\theta^\top (\theta - m_\theta) && \text{(45b)} \\
&= g(m_x, m_\theta) + J_\theta^\top (\theta - m_\theta) && \text{(45c)} \\
&\propto J_\theta^\top \theta\,. && \text{(45d)}
\end{aligned}$$

5.1.3 With respect to θ

$$\begin{aligned}
\mathbb{E}_{q(\theta)}[g(\mathbf{x},\theta)] &= \mathbb{E}_{q(\theta)}[g(m_x, m_\theta) + J_{\mathbf{x}}^\top(\mathbf{x} - m_x) + J_\theta^\top(\theta - m_\theta)] && \text{(46a)}\\
&= g(m_x, m_\theta) + J_{\mathbf{x}}^\top(\mathbf{x} - m_x) + J_\theta^\top \underbrace{\mathbb{E}_{q(\theta)}[(\theta - m_\theta)]}_{0} && \text{(46b)}\\
&= g(m_x, m_\theta) + J_{\mathbf{x}}^\top(\mathbf{x} - m_x) && \text{(46c)}\\
&\propto J_{\mathbf{x}}^\top x\,. && \text{(46d)}
\end{aligned}$$

5.2 Second-order moments

Here things become a bit messier. I will therefore introduce another shorthand: g for $g(m_x, m_\theta)$.

5.2.1 With respect to θ and x

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{x}),q(\theta)}[g(\mathbf{x},\theta)^2] &= \mathbb{E}_{q(\mathbf{x}),q(\theta)}[\big(g + J_{\mathbf{x}}^\top(\mathbf{x} - m_x) + J_\theta^\top(\theta - m_\theta)\big)^2] && \text{(47a)}\\
&= g^2 + 2gJ_{\mathbf{x}}^\top \underbrace{\mathbb{E}_{q(\mathbf{x})}[(\mathbf{x} - m_x)]}_{0} + 2gJ_\theta^\top \underbrace{\mathbb{E}_{q(\theta)}[(\theta - m_\theta)]}_{0}\\
&\quad + 2J_{\mathbf{x}}^\top \underbrace{\mathbb{E}_{q(\mathbf{x})}[(\mathbf{x} - m_x)]}_{0} J_\theta^\top \underbrace{\mathbb{E}_{q(\theta)}[(\theta - m_\theta)]}_{0}\\
&\quad + J_{\mathbf{x}}^\top \mathbb{E}_{q(\mathbf{x})}[(\mathbf{x} - m_x)(\mathbf{x} - m_x)^\top] J_{\mathbf{x}}\\
&\quad + J_\theta^\top \mathbb{E}_{q(\theta)}[(\theta - m_\theta)(\theta - m_\theta)^\top] J_\theta && \text{(47b)}\\
&= g^2 + J_{\mathbf{x}}^\top V_x J_{\mathbf{x}} + J_\theta^\top V_\theta J_\theta\,. && \text{(47c)}
\end{aligned}$$

5.2.2 With respect to x

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{x})}[g(\mathbf{x},\theta)^2] &= \mathbb{E}_{q(\mathbf{x})}[\big(g + J_{\mathbf{x}}^\top(\mathbf{x} - m_x) + J_\theta^\top(\theta - m_\theta)\big)^2] && \text{(48a)}\\
&= g^2 + 2gJ_{\mathbf{x}}^\top \underbrace{\mathbb{E}_{q(\mathbf{x})}[(\mathbf{x} - m_x)]}_{0} + 2gJ_\theta^\top(\theta - m_\theta)\\
&\quad + 2J_{\mathbf{x}}^\top \underbrace{\mathbb{E}_{q(\mathbf{x})}[(\mathbf{x} - m_x)]}_{0} J_\theta^\top(\theta - m_\theta)\\
&\quad + J_{\mathbf{x}}^\top \mathbb{E}_{q(\mathbf{x})}[(\mathbf{x} - m_x)(\mathbf{x} - m_x)^\top] J_{\mathbf{x}}\\
&\quad + J_\theta^\top(\theta - m_\theta)(\theta - m_\theta)^\top J_\theta && \text{(48b)}\\
&= g^2 + 2gJ_\theta^\top(\theta - m_\theta) + J_{\mathbf{x}}^\top V_x J_{\mathbf{x}}\\
&\quad + J_\theta^\top(\theta\theta^\top - m_\theta\theta^\top - \theta m_\theta^\top + m_\theta m_\theta^\top) J_\theta && \text{(48c)}\\
&\propto (g - J_\theta^\top m_\theta) J_\theta^\top \theta + \theta^\top J_\theta (g - J_\theta^\top m_\theta) + \theta^\top J_\theta J_\theta^\top \theta\,. && \text{(48d)}
\end{aligned}$$

5.2.3 With respect to θ

$$
\begin{aligned}
\mathbb{E}_{q(\theta)}[g(\mathbf{x},\theta)^2] &= \mathbb{E}_{q(\theta)}[\big(g + J_{\mathbf{x}}^\top(\mathbf{x}-m_x) + J_\theta^\top(\theta - m_\theta)\big)^2] \qquad (49a)\\
&= g^2 + 2gJ_{\mathbf{x}}^\top(\mathbf{x}-m_x) + 2gJ_\theta^\top \underbrace{\mathbb{E}_{q(\theta)}[(\theta-m_\theta)]}_{0} + \\
&\quad 2J_{\mathbf{x}}^\top(\mathbf{x}-m_x)J_\theta^\top \underbrace{\mathbb{E}_{q(\theta)}[(\theta-m_\theta)]}_{0} \\
&\quad + J_{\mathbf{x}}^\top(\mathbf{x}-m_x)(\mathbf{x}-m_x)^\top J_{\mathbf{x}} \\
&\quad + J_\theta^\top \mathbb{E}_{q(\theta)}[(\theta-m_\theta)(\theta-m_\theta)^\top]J_\theta \qquad (49b)\\
&= gg^\top + g(J_{\mathbf{x}}^\top(\mathbf{x}-m_x))^\top + J_{\mathbf{x}}^\top(\mathbf{x}-m_x)g^\top + J_\theta^\top V_\theta J_\theta \\
&\quad + J_{\mathbf{x}}^\top(\mathbf{x}\mathbf{x}^\top - m_x\mathbf{x}^\top - \mathbf{x}m_x^\top + m_x m_x^\top)J_{\mathbf{x}} \qquad (49c)\\
&\propto (g - J_{\mathbf{x}}^\top m_x)J_{\mathbf{x}}^\top\mathbf{x} + \mathbf{x}^\top J_{\mathbf{x}}(g - J_{\mathbf{x}}^\top m_x) + \mathbf{x}^\top J_{\mathbf{x}}J_{\mathbf{x}}^\top\mathbf{x}\,. \qquad (49d)
\end{aligned}
$$

References

[1] Justin Dauwels, Andrew Eckford, Sascha Korl, and Hans-Andrea Loeliger. Expectation maximization as message passing - Part I: Principles and Gaussian messages. *arXiv:0910.2832*, 2009.

[2] Albert Podusenko, Wouter M. Kouw, and Bert de Vries. Online variational message passing in hierarchical autoregressive models. In *IEEE International Symposium on Information Theory*, pages 1343–1348, 2020.