Supplement (ISIT-2020): Online Variational Message Passing in Hierarchical Auroregressive Models

Albert Podusenko, Wouter. M Kouw, Bert de Vries

AR node

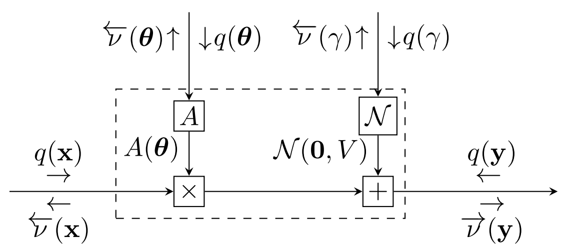

Figure 1: Autoregressive (AR) node.

Derivations

Before proceeding, we need to address the issue of invertability of covariance matrix V with $\epsilon = 0$, i.e.,

$$V = \begin{bmatrix} \frac{1}{\gamma} & 0 & 0 & \dots & 0 \\ 0 & \epsilon & 0 & \dots & \vdots \\ 0 & 0 & \epsilon & \dots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \end{bmatrix}$$

To tackle this problem we assume $\epsilon > 0$, which allows to introduce matrix $W = V^{-1}$ (Appendix D [1]) . In this way, our AR node function can be represented as follows:

$$f(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}, \gamma) = \mathcal{N}\left(\mathbf{y} \mid A(\boldsymbol{\theta})\mathbf{x}, W^{-1}\right)$$

where

$$W = \begin{bmatrix} \gamma & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\epsilon} & 0 & \dots & \vdots \\ 0 & 0 & \frac{1}{\epsilon} & \dots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \end{bmatrix} \tag{1}$$

Note that the introduced precision matrix Eq. (1) enforces AR node function to be a proper distribution [1]. Therefore, vector $\mathbf{y}$ doesn't have merely copied components of vector $\mathbf{x}$, i.e. if $\mathbf{x} = (x_1, x_2, \dots, x_M)$, then

$$\mathbf{y} = \left(\boldsymbol{\theta}^\top \mathbf{x}, x_1 + \mathcal{N}(0, \epsilon), \dots, x_{M-1} + \mathcal{N}(0, \epsilon)\right) = \left(\boldsymbol{\theta}^\top \mathbf{x}, x'_1, \dots, x'_{M-1}\right)$$

The expectation of matrix W:

$$\mathbb{E}_W\left[W\right] = \mathbb{E}_\gamma\left[W\right] = \begin{bmatrix} m_\gamma & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\epsilon} & 0 & \dots & \vdots \\ 0 & 0 & \frac{1}{\epsilon} & \dots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \end{bmatrix}$$

The product $A(\boldsymbol{\theta})\mathbf{x}$ can be separated in the shifting operator $\mathbf{Sx}$ and the inner vector product $\mathbf{c}\mathbf{x}^\top\boldsymbol{\theta}$ in the following way:

$$A(\boldsymbol{\theta})\mathbf{x} = \mathbf{Sx} + \mathbf{c}\mathbf{x}^\top\boldsymbol{\theta} = \underbrace{(\mathbf{S} + \mathbf{c}\boldsymbol{\theta}^\top)}_{A(\boldsymbol{\theta})}\mathbf{x} \tag{2}$$

where

$$\mathbf{S} = \begin{bmatrix} & \mathbf{0}^\top \\ \mathbf{I}_{M-1} & \mathbf{0} \end{bmatrix} \quad \mathbf{c} = (1, 0, \dots, 0)^\top$$

Owing the factorization introduced in Eq. 2 and specific form of matrix W (Eq. 1), our recognition distribution takes the form:

$$q(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}, \gamma) = q(\boldsymbol{\theta})q(\mathbf{x})q(\mathbf{y})q(\gamma) \tag{3}$$

where

$$\begin{aligned} q(\boldsymbol{\theta}) &= \mathcal{N}(m_{\boldsymbol{\theta}}, V_{\boldsymbol{\theta}}) \\ q(\mathbf{x}) &= \mathcal{N}(m_{\mathbf{x}}, V_{\mathbf{x}}) \\ q(\mathbf{y}) &= \mathcal{N}(m_{\mathbf{y}}, V_{\mathbf{y}}) \\ q(\gamma) &= \Gamma(\alpha, \beta) \end{aligned}$$

$\mathcal{N}(m, V)$ denote a Gaussian distribution with mean vector m and covariance matrix V, while $\Gamma(\alpha, \beta)$ - gamma distribution with shape-rate parametrization. For the sake of brevity, in the following paragraphs, we replace $A(\boldsymbol{\theta})$ and $f(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}, \gamma)$ with $\mathbf{A}$ and f respectively.

Update of message to $\mathbf{y}$

$$\begin{aligned}\log \overrightarrow{\nu}(\mathbf{y}) = \mathbb{E}_{\boldsymbol{\theta},\mathbf{x},\gamma} \log f + const = -\frac{1}{2}\, \mathbb{E}_{\boldsymbol{\theta},\mathbf{x},\gamma} \left[(\mathbf{y}-\mathbf{A}\mathbf{x})^{\mathrm{T}} W (\mathbf{y}-\mathbf{A}\mathbf{x})\right] \\ + const \\ = -\frac{1}{2}\, \mathbb{E}_{\boldsymbol{\theta},\mathbf{x},\gamma} \left[\mathbf{y}^\top W \mathbf{y} - (\mathbf{A}\mathbf{x})^\top W \mathbf{y} - \mathbf{y}^\top W \mathbf{A}\mathbf{x} + (\mathbf{A}\mathbf{x})^\top W \mathbf{A}\mathbf{x}\right] \\ + const\end{aligned}$$

As term $(\mathbf{A}\mathbf{x})^\top W \mathbf{A}\mathbf{x}$ doesn't depend on $\mathbf{y}$, we move it into the *const*. Hence,

$$\begin{aligned}\log \overrightarrow{\nu}(\mathbf{y}) &= -\frac{1}{2}\left[\mathbf{y}^\top m_W \mathbf{y} - (\mathbf{S} m_{\mathbf{x}} + \mathbf{c} m_{\mathbf{x}}^\top m_{\boldsymbol{\theta}})^\top m_W \mathbf{y} - \mathbf{y}^\top m_W \underbrace{(\mathbf{S} m_{\mathbf{x}} + \mathbf{c} m_{\mathbf{x}}^\top m_{\boldsymbol{\theta}})}_{m_{\boldsymbol{\theta}} m_{\mathbf{x}}}\right] \\ &\quad + const \\ &= -\frac{1}{2}\left[\mathbf{y}^\top m_W \mathbf{y} - \underbrace{(m_{\boldsymbol{\theta}} m_{\mathbf{x}})^\top m_W}_{\mathbf{z}^\top} \mathbf{y} - \mathbf{y}^\top \underbrace{m_W m_{\boldsymbol{\theta}} m_{\mathbf{x}}}_{\mathbf{z}}\right] + const \\ &= -\frac{1}{2}\left[\mathbf{y}^\top m_W \mathbf{y} - \mathbf{z}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{z} + \mathbf{z}^\top m_W^{-1} \mathbf{z}\right] + const \\ &= -\frac{1}{2}\left[\mathbf{y}^\top m_W \mathbf{y} - \mathbf{z}^\top m_W^{-1} m_W \mathbf{y} - \mathbf{y}^\top m_W m_W^{-1} \mathbf{z} + \mathbf{z}^\top m_W^{-1} m_W m_W^{-1} \mathbf{z}\right] \\ &\quad + const \\ &= -\frac{1}{2}\left[(\mathbf{y} - m_W^{-1}\mathbf{z})^\top m_W (\mathbf{y} - m_W^{-1}\mathbf{z})\right] + const \\ &= -\frac{1}{2}\left[(\mathbf{y} - \underbrace{m_W^{-1} m_W}_{\mathbf{I}} m_{\boldsymbol{\theta}} m_{\mathbf{x}})^\top m_W (\mathbf{y} - m_W^{-1} m_W m_{\boldsymbol{\theta}} m_{\mathbf{x}})\right] + const \\ &= -\frac{1}{2}\left[(\mathbf{y} - m_{\boldsymbol{\theta}} m_{\mathbf{x}})^\top m_W (\mathbf{y} - m_{\boldsymbol{\theta}} m_{\mathbf{x}})\right] + const\end{aligned}$$

Which yields

$$\overrightarrow{\nu}(\mathbf{y}) \propto \mathcal{N}(m_{\boldsymbol{\theta}} m_{\mathbf{x}}, m_W^{-1})$$

Update of message to $\mathbf{x}$

$$\begin{aligned}\log \overleftarrow{\nu}(\mathbf{x}) &= \mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma} \log f + const \\ &= -\frac{1}{2}\, \mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma} \left[\mathbf{y}^{\mathrm{T}} W \mathbf{y} - (\mathbf{A}\mathbf{x})^{\mathrm{T}} W \mathbf{y} - \mathbf{y}^{\mathrm{T}} W \mathbf{A}\mathbf{x} + (\mathbf{A}\mathbf{x})^{\mathrm{T}} W \mathbf{A}\mathbf{x}\right] \\ &\quad + const \\ &= -\frac{1}{2}\, \mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma} \left[-\mathbf{x}^\top \mathbf{A}^\top W \mathbf{y} - \mathbf{y}^\top W \mathbf{A}\mathbf{x} + \mathbf{x}^\top \mathbf{A}^\top W \mathbf{A}\mathbf{x}\right] \\ &\quad + const\end{aligned}$$

We split the expression under the expectation into three terms:
I: $\mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma}\left[-\mathbf{x}^\top \mathbf{A}^\top W\mathbf{y}\right]$
II: $\mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma}\left[-\mathbf{y}^\top W\mathbf{A}\mathbf{x}\right]$ and
III: $\mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma}\left[\mathbf{x}^\top \mathbf{A}^\top W\mathbf{A}\mathbf{x}\right]$
Term I:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma}\left[-\mathbf{x}^\top \mathbf{A}^\top W\mathbf{y}\right] &= -\mathbf{x}^\top \mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma}\left[(\mathbf{S}+\mathbf{c}\boldsymbol{\theta}^\top)^\top W\mathbf{y}\right] \\ &= -\mathbf{x}^\top m_{\mathbf{A}}^\top m_W m_{\mathbf{y}} = -(m_{\mathbf{A}}\mathbf{x})^\top m_W m_{\mathbf{y}}\end{aligned}$$

Term II:

$$\mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma}\left[-\mathbf{y}^\top W\,\mathbf{A}\mathbf{x}\right] = -m_{\mathbf{y}}^\top m_W m_{\mathbf{A}}\mathbf{x}$$

Term III:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma}\left[\mathbf{x}^\top \mathbf{A}^\top W\mathbf{A}\mathbf{x}\right] &= \mathbb{E}_{\boldsymbol{\theta},\gamma}\left[((\mathbf{S}\mathbf{x}+\mathbf{c}\mathbf{x}^\top\boldsymbol{\theta})^\top W(\mathbf{S}\mathbf{x}+\mathbf{c}\mathbf{x}^\top\boldsymbol{\theta})\right] \\ &= \mathbb{E}_{\boldsymbol{\theta}}\left[(\mathbf{c}\mathbf{x}^\top\boldsymbol{\theta})^\top m_W\mathbf{S}\mathbf{x} + (\mathbf{c}\mathbf{x}^\top\boldsymbol{\theta})^\top m_W\mathbf{c}\mathbf{x}^\top\boldsymbol{\theta}\right] \\ &\quad + \mathbb{E}_{\boldsymbol{\theta}}\left[(\mathbf{S}\mathbf{x})^\top m_W\mathbf{c}\mathbf{x}^\top\boldsymbol{\theta} + (\mathbf{S}\mathbf{x})^\top m_W\mathbf{S}\mathbf{x}\right] \\ &= (\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}})^\top m_W\mathbf{S}\mathbf{x} + \mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}^\top\mathbf{x}\mathbf{c}^\top m_W\mathbf{c}\mathbf{x}^\top\boldsymbol{\theta}\right] \\ &\quad + (\mathbf{S}\mathbf{x})^\top m_W\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}} + (\mathbf{S}\mathbf{x})^\top m_W\mathbf{S}\mathbf{x}\end{aligned}$$

where

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}^\top\mathbf{x}\mathbf{c}^\top m_W\mathbf{c}\mathbf{x}^\top\boldsymbol{\theta}\right] = \mathrm{tr}(\mathbf{x}\mathbf{c}^\top m_W\mathbf{c}\mathbf{x}^\top V_{\boldsymbol{\theta}}) + m_{\boldsymbol{\theta}}^\top\mathbf{x}\mathbf{c}^\top m_W\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}}$$

Hence,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta},\mathbf{y},\gamma}\left[\mathbf{x}^\top \mathbf{A}^\top W\mathbf{A}\mathbf{x}\right] \\ &= (\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}})^\top m_W\mathbf{S}\mathbf{x} + \mathrm{tr}(\mathbf{x}\mathbf{c}^\top m_W\mathbf{c}\mathbf{x}^\top V_{\boldsymbol{\theta}}) + m_{\boldsymbol{\theta}}^\top\mathbf{x}\mathbf{c}^\top m_W\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}} \\ &\quad + (\mathbf{S}\mathbf{x})^\top m_W\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}} + (\mathbf{S}\mathbf{x})^\top m_W\mathbf{S}\mathbf{x} \\ &= (\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}})^\top m_W\mathbf{S}\mathbf{x} + \mathbf{x}^\top V_{\boldsymbol{\theta}}\mathbf{c}^\top m_W\mathbf{c}\mathbf{x} + m_{\boldsymbol{\theta}}^\top\mathbf{x}\mathbf{c}^\top m_W\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}} \\ &\quad + (\mathbf{S}\mathbf{x})^\top m_W\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}} + (\mathbf{S}\mathbf{x})^\top m_W\mathbf{S}\mathbf{x} \\ &= (\underbrace{\mathbf{S}\mathbf{x}+\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}}}_{m_{\mathbf{A}}\mathbf{x}})^\top m_W\mathbf{S}\mathbf{x} + (\mathbf{S}\mathbf{x}+\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}})^\top m_W\mathbf{c}\mathbf{x}^\top m_{\boldsymbol{\theta}} \\ &\quad + \mathbf{x}^\top V_{\boldsymbol{\theta}}\underbrace{\mathbf{c}^\top m_W\mathbf{c}}_{m_\gamma}\mathbf{x} \\ &= (m_{\mathbf{A}}\mathbf{x})^\top m_W m_{\mathbf{A}}\mathbf{x} + \mathbf{x}^\top V_{\boldsymbol{\theta}} m_\gamma\mathbf{x}\end{aligned}$$

Putting terms I, II and III together

$$
\begin{aligned}
&\log \overleftarrow{\nu}(\mathbf{x}) \\
&= -\frac{1}{2}\left[-(m_{\mathbf{A}}\mathbf{x})^\top m_W m_{\mathbf{y}} - m_{\mathbf{y}}^\top m_W m_{\mathbf{A}}\mathbf{x} + (m_{\mathbf{A}}\mathbf{x})^\top m_W m_{\mathbf{A}}\mathbf{x} + \mathbf{x}^\top V_{\boldsymbol{\theta}} m_\gamma \mathbf{x}\right] \\
&\quad + const \\
&= -\frac{1}{2}\left[-\mathbf{x}^\top \underbrace{m_{\mathbf{A}}^\top m_W m_{\mathbf{y}}}_{\mathbf{z}} - \underbrace{m_{\mathbf{y}}^\top m_W m_{\mathbf{A}}}_{\mathbf{z}^\top}\mathbf{x} + \mathbf{x}^\top \underbrace{(m_{\mathbf{A}}^\top m_W m_{\mathbf{A}} + V_{\boldsymbol{\theta}} m_\gamma)}_{\mathbf{D}}\mathbf{x}\right] + const \\
&= -\frac{1}{2}\left[\mathbf{x}^\top\mathbf{D}\mathbf{x} - \mathbf{x}^\top\mathbf{z} - \mathbf{z}^\top\mathbf{x} + \mathbf{z}^\top\mathbf{D}^{-\mathbf{1}}\mathbf{z}\right] + const \\
&= -\frac{1}{2}\left[\mathbf{x}^\top\mathbf{D}\mathbf{x} - \mathbf{x}^\top\mathbf{D}\mathbf{D}^{\text{-}1}\mathbf{z} - \mathbf{z}^\top\mathbf{D}^{\text{-}1}\mathbf{D}\mathbf{x} + \mathbf{z}^\top\mathbf{D}^{-\mathbf{1}}\mathbf{D}\mathbf{D}^{\text{-}1}\mathbf{z}\right] + const \\
&= -\frac{1}{2}\left[(\mathbf{x} - \mathbf{D}^{\text{-}1}\mathbf{z})^\top\mathbf{D}(\mathbf{x} - \mathbf{D}^{\text{-}1}\mathbf{z})\right] + const
\end{aligned}
$$

Therefore,

$$\overleftarrow{\nu}(\mathbf{x}) \propto \mathcal{N}(\mathbf{D}^{\text{-}1}\mathbf{z}, \mathbf{D}^{\text{-}1})$$

where

$$
\begin{aligned}
\mathbf{D} &= m_{\mathbf{A}}^\top m_W m_{\mathbf{A}} + V_{\boldsymbol{\theta}} m_\gamma \\
\mathbf{z} &= m_{\mathbf{A}}^\top m_W m_{\mathbf{y}}
\end{aligned}
$$

Update of message to $\boldsymbol{\theta}$

$$
\begin{aligned}
\log \overleftarrow{\nu}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{x},\mathbf{y},\gamma} \log f + const \\
&= -\frac{1}{2}\,\mathbb{E}_{\mathbf{x},\mathbf{y},\gamma}\left[\mathbf{y}^{\mathrm{T}} W \mathbf{y} - (\mathbf{A}\mathbf{x})^{\mathrm{T}} W \mathbf{y} - \mathbf{y}^{\mathrm{T}} W \mathbf{A}\mathbf{x} + (\mathbf{A}\mathbf{x})^{\mathrm{T}} W \mathbf{A}\mathbf{x}\right] \\
&\quad + const \\
&= -\frac{1}{2}\,\mathbb{E}_{\mathbf{x},\mathbf{y},\gamma}\left[-(\mathbf{A}\mathbf{x})^\top W \mathbf{y} - \mathbf{y}^\top W \mathbf{A}\mathbf{x} + (\mathbf{A}\mathbf{x})^\top W \mathbf{A}\mathbf{x}\right] + const \\
&= -\frac{1}{2}\left[-m_{\mathbf{x}}^\top\, \mathbf{A}^\top\, m_W m_{\mathbf{y}} - m_{\mathbf{y}}^\top m_W\, \mathbf{A}\, m_{\mathbf{x}}\right] \\
&\quad - \frac{1}{2}\left[\mathrm{tr}(\mathbf{A}^\top\, m_W\, \mathbf{A}\, V_{\mathbf{x}}) + (\mathbf{A}\, m_{\mathbf{x}})^\top m_W\, \mathbf{A}\, m_{\mathbf{x}}\right] \\
&\quad + const
\end{aligned}
$$

We seek for a quadratic form of vector $\boldsymbol{\theta}$, therefore recalling Eq. 2, we obtain:

$$\begin{aligned}\log \overleftarrow{\nu}(\boldsymbol{\theta}) = &-\frac{1}{2}\left[-(\mathbf{S}m_{\mathbf{x}} + \mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})^{\top} m_W m_{\mathbf{y}} - m_{\mathbf{y}}^{\top} m_W (\mathbf{S}m_{\mathbf{x}} + \mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})\right] \\ &- \frac{1}{2}\left[\mathrm{tr}\left[m_W(\mathbf{S} + \mathbf{c}\boldsymbol{\theta}^{\top})V_{\mathbf{x}}(\mathbf{S} + \mathbf{c}\boldsymbol{\theta}^{\top})^{\top}\right]\right] \\ &- \frac{1}{2}\left[(\mathbf{S}m_{\mathbf{x}} - \mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})^{\top} m_W (\mathbf{S}m_{\mathbf{x}} - \mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})\right] \\ &+ const\end{aligned}$$

Let us work out the term $\mathrm{tr}\left[m_W(\mathbf{S} + \mathbf{c}\boldsymbol{\theta}^{\top})V_{\mathbf{x}}(\mathbf{S} + \mathbf{c}\boldsymbol{\theta}^{\top})^{\top}\right]$ separately.

$$\begin{aligned}&\mathrm{tr}\left[m_W(\mathbf{S} + \mathbf{c}\boldsymbol{\theta}^{\top})V_{\mathbf{x}}(\mathbf{S} + \mathbf{c}\boldsymbol{\theta}^{\top})^{\top}\right] \\ &\quad = \mathrm{tr}[m_W\mathbf{S}V_{\mathbf{x}}\mathbf{S}^{\top} + m_W\mathbf{c}\boldsymbol{\theta}^{\top}V_{\mathbf{x}}\mathbf{S}^{\top} + m_W\mathbf{c}\boldsymbol{\theta}^{\top}V_{\mathbf{x}}\boldsymbol{\theta}\mathbf{c}^{\top} + m_W\mathbf{S}V_{\mathbf{x}}\boldsymbol{\theta}\mathbf{c}^{\top}] \qquad (4) \\ &\quad = \mathrm{tr}(m_W\mathbf{S}V_{\mathbf{x}}\mathbf{S}^{\top}) + \boldsymbol{\theta}^{\top}V_{\mathbf{x}}\mathbf{S}^{\top}m_W\mathbf{c} + \mathbf{c}^{\top}m_W\mathbf{c}\boldsymbol{\theta}^{\top}V_{\mathbf{x}}\boldsymbol{\theta} + \mathbf{c}^{\top}m_W\mathbf{S}V_{\mathbf{x}}\boldsymbol{\theta}\end{aligned}$$

Owing that $\mathbf{S}^{\top}\boldsymbol{\Sigma}\mathbf{c} = \mathbf{0}$ and $\mathbf{c}^{\top}\boldsymbol{\Sigma}\mathbf{S} = \mathbf{0}^{\top}$, where $\boldsymbol{\Sigma}$ is an arbitrary diagonal matrix.

$$\begin{aligned}\log \overleftarrow{\nu}(\boldsymbol{\theta}) &= -\frac{1}{2}\Big[-(\mathbf{S}m_{\mathbf{x}})^{\top}m_W m_{\mathbf{y}} - (\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})^{\top}m_W m_{\mathbf{y}} - m_{\mathbf{y}}^{\top}m_W\mathbf{S}m_{\mathbf{x}} \\ &\quad - m_{\mathbf{y}}^{\top}m_W\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta} + \mathrm{tr}(m_W\mathbf{S}V_{\mathbf{x}}\mathbf{S}^{\top}) + \underbrace{\boldsymbol{\theta}^{\top}V_{\mathbf{x}}\mathbf{S}^{\top}m_W\mathbf{c}}_{0} \\ &\quad + \mathbf{c}^{\top}m_W\mathbf{c}\boldsymbol{\theta}^{\top}V_{\mathbf{x}}\boldsymbol{\theta} + \underbrace{\mathbf{c}^{\top}m_W\mathbf{S}V_{\mathbf{x}}\boldsymbol{\theta}}_{0} \\ &\quad + (\mathbf{S}m_{\mathbf{x}})^{\top}m_W\mathbf{S}m_{\mathbf{x}} - \underbrace{(\mathbf{S}m_{\mathbf{x}})^{\top}m_W\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta}}_{0} \\ &\quad - \underbrace{(\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})^{\top}m_W\mathbf{S}m_{\mathbf{x}}}_{0} + (\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})^{\top}m_W\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta}\Big] + const \\ &= -\frac{1}{2}\Big[-(\mathbf{S}m_{\mathbf{x}})^{\top}m_W m_{\mathbf{y}} - (\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})^{\top}m_W m_{\mathbf{y}} - m_{\mathbf{y}}^{\top}m_W\mathbf{S}m_{\mathbf{x}} \\ &\quad - m_{\mathbf{y}}^{\top}m_W\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta} + \mathrm{tr}(m_W\mathbf{S}V_{\mathbf{x}}\mathbf{S}^{\top}) + \mathbf{c}^{\top}m_W\mathbf{c}\boldsymbol{\theta}^{\top}V_{\mathbf{x}}\boldsymbol{\theta} \\ &\quad + (\mathbf{S}m_{\mathbf{x}})^{\top}m_W\mathbf{S}m_{\mathbf{x}} + (\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})^{\top}m_W\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta}\Big] + const\end{aligned}$$

We move terms which do not depend on $\boldsymbol{\theta}$ to the *const*, hence

$$\begin{aligned}
\log \overleftarrow{\nu}(\boldsymbol{\theta}) &= -\frac{1}{2}\Big[(-(\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})^{\top}m_W m_{\mathbf{y}} - m_{\mathbf{y}}^{\top}m_W\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta} + \mathbf{c}^{\top}m_W\mathbf{c}\boldsymbol{\theta}^{\top}V_{\mathbf{x}}\boldsymbol{\theta} \\
&\qquad + (\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta})^{\top}m_W\mathbf{c}m_{\mathbf{x}}^{\top}\boldsymbol{\theta}\Big] + const \\
&= -\frac{1}{2}\Big[\boldsymbol{\theta}^{\top}\underbrace{(V_{\mathbf{x}}\mathbf{c}^{\top}m_W\mathbf{c} + m_{\mathbf{x}}\mathbf{c}^{\top}m_W\mathbf{c}m_{\mathbf{x}}^{\mathrm{T}})}_{\mathbf{D}}\boldsymbol{\theta} - \underbrace{m_{\mathbf{y}}^{\top}m_W\mathbf{c}m_{\mathbf{x}}^{\top}}_{\mathbf{z}^{\top}}\boldsymbol{\theta} \\
&\qquad - \boldsymbol{\theta}^{\top}\underbrace{m_{\mathbf{x}}\mathbf{c}^{\top}m_W m_{\mathbf{y}}}_{\mathbf{z}}\Big] + const \\
&= -\frac{1}{2}\left[\boldsymbol{\theta}^{\top}\mathbf{D}\boldsymbol{\theta} - \mathbf{z}^{\top}\boldsymbol{\theta} - \boldsymbol{\theta}^{\top}\mathbf{z} + \mathbf{z}^{\top}\mathbf{D}^{-1}\mathbf{z}\right] + const \\
&= -\frac{1}{2}\left[(\boldsymbol{\theta}-\mathbf{D}^{-1}\mathbf{z})^{\top}\mathbf{D}(\boldsymbol{\theta}-\mathbf{D}^{-1}\mathbf{z})\right] + const
\end{aligned}$$

Hence,

$$\overleftarrow{\nu}(\boldsymbol{\theta}) \propto \mathcal{N}(\mathbf{D}^{-1}\mathbf{z}, \mathbf{D}^{-1})$$

where

$$\begin{aligned}
&\mathbf{z} = m_{\mathbf{x}}\mathbf{c}^{\top}m_W m_{\mathbf{y}} = m_{\gamma}(m_{\mathbf{x}} \circ m_{\mathbf{y}}) \\
&(\circ)\text{denote Hadamard product} \\
&\mathbf{D} = (V_{\mathbf{x}}\mathbf{c}^{\top}m_W\mathbf{c} + m_{\mathbf{x}}\mathbf{c}^{\top}m_W\mathbf{c}m_{\mathbf{x}}^{\mathrm{T}}) = V_{\mathbf{x}}m_{\gamma} + m_{\mathbf{x}}m_{\gamma}m_{\mathbf{x}}^{\mathrm{T}}
\end{aligned}$$

Update of message to γ

$$\begin{aligned}
\log \overleftarrow{\nu}(\gamma) &= \mathbb{E}_{\mathbf{x},\mathbf{y},\boldsymbol{\theta}} \log f + const \\
&= \log(|W|^{\frac{1}{2}}) - \frac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y},\boldsymbol{\theta}}\left[(\mathbf{y}-\mathbf{A}\mathbf{x})^{\top}W(\mathbf{y}-\mathbf{A}\mathbf{x})\right] + const \\
&= \log(|W|^{\frac{1}{2}}) \\
&\quad - \frac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y},\boldsymbol{\theta}}\left[\mathbf{y}^{\top}W\mathbf{y} - \mathbf{x}^{\top}\mathbf{A}^{\top}W\mathbf{y} - \mathbf{y}^{\top}W\mathbf{A}\mathbf{x} + \mathbf{x}^{\top}\mathbf{A}^{\top}W\mathbf{A}\mathbf{x}\right] \\
&\quad + const \\
&= \log(|W|^{\frac{1}{2}}) - \frac{1}{2}\Big[\mathrm{tr}(WV_{\mathbf{y}}) + m_{\mathbf{y}}^{\top}Wm_{\mathbf{y}} - (m_{\mathbf{A}}m_{\mathbf{x}})^{\top}Wm_{\mathbf{y}} \\
&\quad - m_{\mathbf{y}}^{\top}Wm_{\mathbf{A}}m_{\mathbf{x}} + \mathbb{E}_{\mathbf{x},\boldsymbol{\theta}}[\mathbf{x}^{\top}\mathbf{A}^{\top}W\mathbf{A}\mathbf{x}]\Big] + const
\end{aligned}$$

Let us work out the remaining expectation term separately

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\theta}}[\mathbf{x}^{\top}\mathbf{A}^{\top}W\mathbf{A}\mathbf{x}] = \mathbb{E}_{\boldsymbol{\theta}}[\mathrm{tr}\left(\mathbf{A}^{\top}W\mathbf{A}V_{\mathbf{x}}\right) + (\mathbf{A}m_{\mathbf{x}})^{\top}W\mathbf{A}m_{\mathbf{x}}]$$

where

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}}[\mathrm{tr}\left(\mathbf{A}^\top W\, \mathbf{A}\, V_{\mathbf{x}}\right)] \\
&\quad = \mathrm{tr}\left(W\, \mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{S}+\mathbf{c}\boldsymbol{\theta}^\top)V_{\mathbf{x}}(\mathbf{S}+\mathbf{c}\boldsymbol{\theta}^\top)^\top]\right) \\
&\quad = \mathrm{tr}\left(W\, \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{S}V_{\mathbf{x}}\mathbf{S}^\top + \mathbf{c}\boldsymbol{\theta}^\top V_{\mathbf{x}}\mathbf{S}^\top + \mathbf{c}\boldsymbol{\theta}^\top V_{\mathbf{x}}\boldsymbol{\theta}\mathbf{c}^\top + \mathbf{S}V_{\mathbf{x}}\boldsymbol{\theta}\mathbf{c}^\top\right]\right) \\
&\quad = \mathrm{tr}(W\mathbf{S}V_{\mathbf{x}}\mathbf{S}^\top) + \underbrace{m_{\boldsymbol{\theta}}^\top V_{\mathbf{x}}\mathbf{S}^\top W\mathbf{c}}_{0} + \underbrace{\mathbf{c}^\top W\mathbf{c}}_{\gamma} m_{\boldsymbol{\theta}}^\top V_{\mathbf{x}} m_{\boldsymbol{\theta}} + \underbrace{\mathbf{c}^\top W\mathbf{S}V_{\mathbf{x}} m_{\boldsymbol{\theta}}}_{0} \\
&\quad = \mathrm{tr}(\mathbf{S}^\top W\mathbf{S}V_{\mathbf{x}}) + \gamma m_{\boldsymbol{\theta}}^\top V_{\mathbf{x}} m_{\boldsymbol{\theta}} \qquad (5)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}}\left[(\mathbf{A}\, m_{\mathbf{x}})^\top W\, \mathbf{A}\, m_{\mathbf{x}}\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{tr}\left(m_{\mathbf{x}}^\top\, \mathbf{A}^\top\, W\, \mathbf{A}\, m_{\mathbf{x}}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathrm{tr}\left(W\, \mathbf{A}\, m_{\mathbf{x}} m_{\mathbf{x}}^\top\, \mathbf{A}^\top\right)\right] \\
&= \mathrm{tr}\left(W\, \mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{S}+\mathbf{c}\boldsymbol{\theta}^\top) m_{\mathbf{x}} m_{\mathbf{x}}^\top (\mathbf{S}+\mathbf{c}\boldsymbol{\theta}^\top)^\top]\right) \\
&= \mathrm{tr}(\mathbf{S}^\top W\mathbf{S} m_{\mathbf{x}} m_{\mathbf{x}}^\top) + \gamma m_{\mathbf{x}}^\top V_{\boldsymbol{\theta}} m_{\mathbf{x}} + \gamma m_{\boldsymbol{\theta}}^\top m_{\mathbf{x}} m_{\mathbf{x}}^\top m_{\boldsymbol{\theta}} \quad (6)
\end{aligned}$$

Before proceeding, let us focus on $\mathbf{S}^\top W\mathbf{S}$:

$$\mathbf{S}^\top W\mathbf{S} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{M-1} \\ & \mathbf{0}^\top \end{bmatrix} \begin{bmatrix} \gamma & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\epsilon} & 0 & \dots & \vdots \\ 0 & 0 & \frac{1}{\epsilon} & \dots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \end{bmatrix} \begin{bmatrix} \mathbf{0}^\top & \\ \mathbf{I}_{M-1} & \mathbf{0} \end{bmatrix}$$

Obviously, there will be no entry of γ in the resulting matrix. Therefore the terms which incorporate $\mathbf{S}^\top W\mathbf{S}$ can be moved to the *const*.

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\theta}}[\mathbf{x}^\top\, \mathbf{A}^\top\, W\, \mathbf{A}\mathbf{x}] = \gamma m_{\boldsymbol{\theta}}^\top V_{\mathbf{x}} m_{\boldsymbol{\theta}} + \gamma m_{\mathbf{x}}^\top V_{\boldsymbol{\theta}} m_{\mathbf{x}} + \gamma m_{\boldsymbol{\theta}}^\top m_{\mathbf{x}} m_{\mathbf{x}}^\top m_{\boldsymbol{\theta}} + const$$

This leads to

$$\begin{aligned}
\log \overleftarrow{\nu}(\gamma) = \log(|W|^{\frac{1}{2}}) - \frac{1}{2}\Big[&\mathrm{tr}(WV_{\mathbf{y}}) + m_{\mathbf{y}}^\top W m_{\mathbf{y}} - (m_{\mathbf{A}} m_{\mathbf{x}})^\top W m_{\mathbf{y}} \\
&- m_{\mathbf{y}}^\top W m_{\mathbf{A}} m_{\mathbf{x}} + \gamma m_{\boldsymbol{\theta}}^\top V_{\mathbf{x}} m_{\boldsymbol{\theta}} + \gamma m_{\mathbf{x}}^\top V_{\boldsymbol{\theta}} m_{\mathbf{x}} + \gamma m_{\boldsymbol{\theta}}^\top m_{\mathbf{x}} m_{\mathbf{x}}^\top m_{\boldsymbol{\theta}}\Big] \\
&+ const
\end{aligned}$$

As the resulting message should depend solely on γ we need to get rid of all terms which incorporate matrix W. Let us consider these terms separately:
I: $\log(|W|^{\frac{1}{2}})$
II: $\mathrm{tr}(WV_{\mathbf{y}})$
III: $m_{\mathbf{y}}^\top W m_{\mathbf{y}}$
IV: $(m_{\mathbf{A}} m_{\mathbf{x}})^\top W m_{\mathbf{y}}$

V: $m_{\mathbf{y}}^\top W m_{\mathbf{A}} m_{\mathbf{x}}$

Term I:

$$\log(|W|^{\frac{1}{2}}) = \frac{1}{2}\log\begin{vmatrix} \gamma & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\epsilon} & 0 & \dots & \vdots \\ 0 & 0 & \frac{1}{\epsilon} & \dots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \end{vmatrix}$$
$$= \frac{1}{2}\log\gamma + \frac{1}{2}(1-M)\log(\epsilon) = \log\gamma^{\frac{1}{2}} + const$$

Term II:

$$\operatorname{tr}(WV_{\mathbf{y}}) = \operatorname{tr}\left(\begin{bmatrix} \gamma & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\epsilon} & 0 & \dots & \vdots \\ 0 & 0 & \frac{1}{\epsilon} & \dots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \end{bmatrix} V_{\mathbf{y}}\right)$$
$$= \gamma V_{\mathbf{y}}^{(1,1)} + const$$

where $V_{\mathbf{y}}^{(1,1)}$ is the first element of matrix $V_{\mathbf{y}}$.

Term III:

$$m_{\mathbf{y}}^\top W m_{\mathbf{y}} = \gamma m_{\mathbf{y}}^{(1)} m_{\mathbf{y}}^{(1)} + const$$

where $m_{\mathbf{y}}^{(1)}$ is the first element of $m_{\mathbf{y}}$.

Both terms IV and V result into the same scalar, i.e.

$$m_{\mathbf{y}}^\top W m_{\mathbf{A}} m_{\mathbf{x}} = (m_{\mathbf{A}} m_{\mathbf{x}})^\top W m_{\mathbf{y}}$$
$$= \left(m_{\boldsymbol{\theta}}^\top m_{\mathbf{x}}, m_{\mathbf{x}}^{(1)}, \dots, m_{\mathbf{x}}^{(M-1)}\right) \begin{bmatrix} \gamma & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\epsilon} & 0 & \dots & \vdots \\ 0 & 0 & \frac{1}{\epsilon} & \dots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \end{bmatrix} \begin{bmatrix} m_{\mathbf{y}}^{(1)} \\ m_{\mathbf{y}}^{(2)} \\ \vdots \\ m_{\mathbf{y}}^{(M)} \end{bmatrix}$$
$$= m_{\boldsymbol{\theta}}^\top m_{\mathbf{x}} \gamma m_{\mathbf{y}}^{(1)} + \sum_{i=1}^{M-1} \frac{m_{\mathbf{x}}^{(i)}}{\epsilon} m_{\mathbf{y}}^{(i+1)} = \gamma m_{\mathbf{y}}^{(1)} m_{\boldsymbol{\theta}}^\top m_{\mathbf{x}} + const$$

Hence

$$
\begin{aligned}
\log \overleftarrow{\nu}(\gamma) &= \log \gamma^{\frac{1}{2}} - \frac{1}{2}\Big(\gamma V_{\mathbf{y}}^{(1,1)} + \gamma m_{\mathbf{y}}^{(1)} m_{\mathbf{y}}^{(1)} - 2\gamma m_{\mathbf{y}}^{(1)} m_{\boldsymbol{\theta}}^{\top} m_{\mathbf{x}} \\
&\qquad + \gamma m_{\boldsymbol{\theta}}^{\top} V_{\mathbf{x}} m_{\boldsymbol{\theta}} + \gamma m_{\mathbf{x}}^{\top} V_{\boldsymbol{\theta}} m_{\mathbf{x}} + \gamma m_{\boldsymbol{\theta}}^{\top} m_{\mathbf{x}} m_{\mathbf{x}}^{\top} m_{\boldsymbol{\theta}}\Big) + const \\
&= \log \gamma^{\frac{1}{2}} - \frac{\gamma}{2}\Big(V_{\mathbf{y}}^{(1,1)} + m_{\mathbf{y}}^{(1)} m_{\mathbf{y}}^{(1)} - 2 m_{\mathbf{y}}^{(1)} m_{\boldsymbol{\theta}}^{\top} m_{\mathbf{x}} \\
&\qquad + m_{\boldsymbol{\theta}}^{\top} V_{\mathbf{x}} m_{\boldsymbol{\theta}} + m_{\mathbf{x}}^{\top} V_{\boldsymbol{\theta}} m_{\mathbf{x}} + m_{\boldsymbol{\theta}}^{\top} m_{\mathbf{x}} m_{\mathbf{x}}^{\top} m_{\boldsymbol{\theta}}\Big) + const
\end{aligned}
$$

After exponentiating $\log \overleftarrow{\nu}(\gamma)$ it yields a gamma distribution:

$$
\overleftarrow{\nu}(\gamma) \propto \gamma^{\frac{1}{2}} \exp\left\{-\frac{\gamma}{2} B\right\}
$$

or

$$
\overleftarrow{\nu}(\gamma) \propto \Gamma\left(\frac{3}{2}, \frac{B}{2}\right)
$$

where

$$
\begin{aligned}
B = {} & V_{\mathbf{y}}^{(1,1)} + m_{\mathbf{y}}^{(1)} m_{\mathbf{y}}^{(1)} - 2 m_{\mathbf{y}}^{(1)} m_{\boldsymbol{\theta}}^{\top} m_{\mathbf{x}} + m_{\mathbf{x}}^{\top} V_{\boldsymbol{\theta}} m_{\mathbf{x}} \\
& + m_{\boldsymbol{\theta}}^{\top} (V_{\mathbf{x}} + m_{\mathbf{x}} m_{\mathbf{x}}^{\top}) m_{\boldsymbol{\theta}}
\end{aligned}
$$

References

[1] J. Dauwels, A. Eckford, S. Korl, and H.-A. Loeliger, "Expectation maximization as message passing-part I: Principles and gaussian messages," *arXiv preprint arXiv:0910.2832*, 2009. [Online]. Available: http://arxiv.org/abs/0910.2832