PROBABILISTIC FILTER AND SMOOTHER FOR VARIATIONAL INFERENCE OF BAYESIAN LINEAR DYNAMICAL SYSTEMS - APPENDIX

Julian Neri[*] *Roland Badeau*[†] *Philippe Depalle*[*]

[*]McGill University, CIRMMT, Montréal, Canada.
[†]LTCI, Télécom Paris, Institut Polytechnique de Paris, France.

Contents

[*]Thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN- 2018-05662) for funding.

A. FORWARD PASS (FILTERING) DERIVATION

The forward pass calculates the statistics of the marginal distribution

$$q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n}) = \frac{p(\mathbf{y}_n|\mathbf{x}_n)q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1})}{q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1})} \tag{1}$$

$$= \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_n, \mathbf{V}_n). \tag{2}$$

The predictive distribution $q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1})$, marginal likelihood $q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1})$, and marginal posterior $q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n})$ are derived in turn. These derivations involve quadratic forms of random variables, as detailed in Section D, and Schur complements, provided in Section E.

A.1. Predictive distribution

First, we evaluate the predictive distribution and show that it is Gaussian distributed with mean $\mathbf{m}_{n-1}$ and covariance $\mathbf{P}_{n-1}$. We retrieve the predictive distribution by marginalizing out $\mathbf{x}_{n-1}$ from the transition probability weighted by the previous marginal:

$$q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1}) = \int p(\mathbf{x}_n|\mathbf{x}_{n-1})q_{\mathbf{x}}(\mathbf{x}_{n-1}|\mathbf{y}_{1:n-1})d\mathbf{x}_{n-1} \tag{3}$$

$$= \int \mathcal{N}(\mathbf{x}_n|\mathbf{A}\mathbf{x}_{n-1} + \mathbf{B}\mathbf{u}_n, \mathbf{Q})\mathcal{N}(\mathbf{x}_{n-1}|\boldsymbol{\mu}_{n-1}, \mathbf{V}_{n-1})d\mathbf{x}_{n-1} \tag{4}$$

$$= \mathcal{N}(\mathbf{x}_n|\mathbf{m}_{n-1}, \mathbf{P}_{n-1}) \tag{5}$$

Collecting the quadratic terms over $\mathbf{x}_n$ and $\mathbf{x}_{n-1}$, we get

$$\boldsymbol{\Omega}_{11} = \mathbf{V}_{n-1}^{-1} + \langle \mathbf{A}^\mathsf{T}\mathbf{Q}^{-1}\mathbf{A}\rangle \tag{6}$$

$$= \mathbf{V}_{n-1}^{-1} + \boldsymbol{\Sigma}_{AQA} + \langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle \tag{7}$$

$$\boldsymbol{\Omega}_{12} = \langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle \tag{8}$$

$$\boldsymbol{\Omega}_{21} = \langle \mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle \tag{9}$$

$$\boldsymbol{\Omega}_{22} = \langle \mathbf{Q}^{-1}\rangle \tag{10}$$

Using the Schur complement (refer to Section E for complements) gives the following analytic inverse

$$\mathbf{P}_{n-1} = \big(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\big)^{-1} \tag{11}$$

$$= \boldsymbol{\Omega}_{22}^{-1} + \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\mathbf{F}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1} \tag{12}$$

where

$$\mathbf{F}_{11}^{-1} = \big(\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\big)^{-1}. \tag{13}$$

Substituting in and simplifying gives

$$\mathbf{F}_{11}^{-1} = \big(\mathbf{V}_{n-1}^{-1} + \boldsymbol{\Sigma}_{AQA} + \langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle - \langle \mathbf{A}^\mathsf{T}\rangle\cancel{\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{Q}^{-1}\rangle^{-1}}\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle\big)^{-1} \tag{14}$$

$$= \big(\mathbf{V}_{n-1}^{-1} + \boldsymbol{\Sigma}_{AQA} + \cancel{\langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle} - \cancel{\langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle}\big)^{-1} \tag{15}$$

$$= \big(\mathbf{V}_{n-1}^{-1} + \boldsymbol{\Sigma}_{AQA}\big)^{-1}. \tag{16}$$

It is computationally costly and numerically problematic to invert $\mathbf{V}_{n-1}^{-1}$ at every time n. We apply the Woodbury matrix identity to retrieve the following preferred analytic expression for $\mathbf{F}_{11}^{-1}$,

$$\mathbf{F}_{11}^{-1} = \big(\mathbf{V}_{n-1}^{-1} + \boldsymbol{\Sigma}_{AQA}\big)^{-1} \tag{17}$$

$$= \mathbf{V}_{n-1} - \mathbf{V}_{n-1}\big(\mathbf{I} + \boldsymbol{\Sigma}_{AQA}\mathbf{V}_{n-1}\big)^{-1}\boldsymbol{\Sigma}_{AQA}\mathbf{V}_{n-1} \tag{18}$$

$$= \big(\mathbf{I} - \mathbf{V}_{n-1}\big(\mathbf{I} + \boldsymbol{\Sigma}_{AQA}\mathbf{V}_{n-1}\big)^{-1}\boldsymbol{\Sigma}_{AQA}\big)\mathbf{V}_{n-1} \tag{19}$$

$$= \mathbf{G}_{n-1}\mathbf{V}_{n-1} \tag{20}$$

where we have defined

$$\mathbf{G}_{n-1} = \mathbf{I} - \mathbf{V}_{n-1}\big(\mathbf{I} + \mathbf{\Sigma}_{AQA}\mathbf{V}_{n-1}\big)^{-1}\mathbf{\Sigma}_{AQA}. \tag{21}$$

Substituting this result back into the expression for the covariance $\mathbf{P}_{n-1}$ and simplifying gives

$$\begin{aligned}
\mathbf{P}_{n-1} &= \mathbf{\Omega}_{22}^{-1} + \mathbf{\Omega}_{22}^{-1}\mathbf{\Omega}_{21}\mathbf{F}_{11}^{-1}\mathbf{\Omega}_{12}\mathbf{\Omega}_{22}^{-1} && (22)\\
&= \langle\mathbf{Q}^{-1}\rangle^{-1} + \cancel{\langle\mathbf{Q}^{-1}\rangle^{-1}\langle\mathbf{Q}^{-1}\rangle}\langle\mathbf{A}\rangle\mathbf{G})n - 1\mathbf{V}_{n-1}\langle\mathbf{A}^{\mathsf{T}}\rangle\cancel{\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{Q}^{-1}\rangle^{-1}} && (23)\\
&= \langle\mathbf{Q}\rangle + \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}\langle\mathbf{A}^{\mathsf{T}}\rangle. && (24)
\end{aligned}$$

Next, we derive the expression for the mean $\mathbf{m}_{n-1}$. Completing the square for $\mathbf{x}_n$ and $\mathbf{x}_{n-1}$, the linear terms over $\mathbf{x}_{n-1}$ and $\mathbf{x}_n$ are as follows.

$$\begin{aligned}
\boldsymbol{\ell}_1 &= \mathbf{V}_{n-1}^{-1}\boldsymbol{\mu}_{n-1} - \langle\mathbf{A}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{B}\rangle\mathbf{u}_n && (25)\\
&= \mathbf{V}_{n-1}^{-1}\boldsymbol{\mu}_{n-1} - \big(\langle\mathbf{A}^{\mathsf{T}}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{B}\rangle + \mathbf{\Sigma}_{AQB}\big)\mathbf{u}_n && (26)\\
\boldsymbol{\ell}_2 &= \langle\mathbf{Q}^{-1}\rangle\langle\mathbf{B}\rangle\mathbf{u}_n && (27)
\end{aligned}$$

The mean is then given by

$$\mathbf{m}_{n-1} = \mathbf{P}_{n-1}\big(\boldsymbol{\ell}_2 - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\boldsymbol{\ell}_1\big). \tag{28}$$

Using the Schur complement, we have

$$\begin{aligned}
\mathbf{P}_{n-1}\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} &= \mathbf{F}_{22}^{-1}\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} && (29)\\
&= \mathbf{\Omega}_{22}^{-1}\mathbf{\Omega}_{21}\mathbf{F}_{11}^{-1} && (30)\\
&= \cancel{\langle\mathbf{Q}^{-1}\rangle^{-1}\langle\mathbf{Q}^{-1}\rangle}\langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1} && (31)\\
&= \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}. && (32)
\end{aligned}$$

Substituting this term back into the expression for the mean and simplifying gives the following.

$$\begin{aligned}
\mathbf{m}_{n-1} &= \mathbf{P}_{n-1}\big(\boldsymbol{\ell}_2 + \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\boldsymbol{\ell}_1\big) && (33)\\
&= \mathbf{P}_{n-1}\boldsymbol{\ell}_2 + \mathbf{P}_{n-1}\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\boldsymbol{\ell}_1 && (34)\\
&= \big(\langle\mathbf{Q}\rangle + \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}\langle\mathbf{A}^{\mathsf{T}}\rangle\big)\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{B}\rangle\mathbf{u}_n && (35)\\
&\quad + \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}\big(\mathbf{V}_{n-1}^{-1}\boldsymbol{\mu}_{n-1} - \big(\langle\mathbf{A}^{\mathsf{T}}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{B}\rangle + \mathbf{\Sigma}_{AQB}\big)\mathbf{u}_n\big) && (36)\\
&= \cancel{\langle\mathbf{Q}\rangle\langle\mathbf{Q}^{-1}\rangle}\langle\mathbf{B}\rangle\mathbf{u}_n + \cancel{\langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}\langle\mathbf{A}^{\mathsf{T}}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{B}\rangle\mathbf{u}_n} && (37)\\
&\quad + \cancel{\langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}\mathbf{V}_{n-1}^{-1}}\boldsymbol{\mu}_{n-1} - \cancel{\langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}\langle\mathbf{A}^{\mathsf{T}}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{B}\rangle\mathbf{u}_n} - \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}\mathbf{\Sigma}_{AQB}\mathbf{u}_n && (38)\\
&= \langle\mathbf{B}\rangle\mathbf{u}_n + \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\boldsymbol{\mu}_{n-1} - \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}\mathbf{\Sigma}_{AQB}\mathbf{u}_n && (39)
\end{aligned}$$

In summary, the marginalizing over $\mathbf{x}_{n-1}$ gives the predictive distribution

$$q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1}) = \mathcal{N}(\mathbf{x}_n|\mathbf{m}_{n-1}, \mathbf{P}_{n-1}) \tag{40}$$

where

$$\begin{aligned}
\mathbf{G}_{n-1} &= \mathbf{I} - \mathbf{V}_{n-1}\big(\mathbf{I} + \mathbf{\Sigma}_{AQA}\mathbf{V}_{n-1}\big)^{-1}\mathbf{\Sigma}_{AQA} && (41)\\
\mathbf{m}_{n-1} &= \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\left(\boldsymbol{\mu}_{n-1} - \mathbf{V}_{n-1}\mathbf{\Sigma}_{AQB}\mathbf{u}_n\right) + \langle\mathbf{B}\rangle\mathbf{u}_n && (42)\\
\mathbf{P}_{n-1} &= \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}\langle\mathbf{A}^{\mathsf{T}}\rangle + \langle\mathbf{Q}\rangle. && (43)
\end{aligned}$$

Initially at time $n = 1$, the predictive distribution is simply equal to the prior over the initial state $p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\mathbf{m}_0, \mathbf{P}_0)$.

A.2. Marginal likelihood

Next, we derive the normalization term for the marginal posterior probability. The distribution is found by taking the integral over (marginalizing out) $\mathbf{x}_n$, resulting in a Gaussian distribution with mean denoted by $\widehat{\mathbf{y}}_n$ and covariance denoted by $\mathbf{S}_n$,

$$q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1}) = \int p(\mathbf{y}_n|\mathbf{x}_n) q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1}) d\mathbf{x}_n \tag{44}$$

$$= \int \mathcal{N}(\mathbf{y}_n|\mathbf{C}\mathbf{x}_n + \mathbf{D}\mathbf{u}_n, \mathbf{R}) \mathcal{N}(\mathbf{x}_n|\mathbf{m}_{n-1}, \mathbf{P}_{n-1}) d\mathbf{x}_n \tag{45}$$

$$= \mathcal{N}(\mathbf{y}|\widehat{\mathbf{y}}_n, \mathbf{S}_n). \tag{46}$$

Collecting the quadratic terms over $\mathbf{x}_n$ and $\mathbf{y}_n$:

$$\boldsymbol{\Omega}_{11} = \mathbf{P}_{n-1}^{-1} + \langle \mathbf{C}^{\mathsf{T}} \mathbf{R}^{-1} \mathbf{C} \rangle \tag{47}$$

$$= \mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC} + \langle \mathbf{C}^{\mathsf{T}} \rangle \langle \mathbf{R}^{-1} \rangle \langle \mathbf{C} \rangle \tag{48}$$

$$\boldsymbol{\Omega}_{12} = \langle \mathbf{C}^{\mathsf{T}} \rangle \langle \mathbf{R}^{-1} \rangle \tag{49}$$

$$\boldsymbol{\Omega}_{21} = \langle \mathbf{R}^{-1} \rangle \langle \mathbf{C} \rangle \tag{50}$$

$$\boldsymbol{\Omega}_{22} = \langle \mathbf{R}^{-1} \rangle. \tag{51}$$

Using the Schur complement, we have

$$\mathbf{S}_n = \big(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21} \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{12}\big)^{-1} \tag{52}$$

$$= \boldsymbol{\Omega}_{22}^{-1} + \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\Omega}_{21} \mathbf{F}_{11}^{-1} \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-1} \tag{53}$$

where

$$\mathbf{F}_{11}^{-1} = \big(\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\Omega}_{21}\big)^{-1} \tag{54}$$

$$= \big(\mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC} + \langle \mathbf{C}^{\mathsf{T}} \rangle \langle \mathbf{R}^{-1} \rangle \langle \mathbf{C} \rangle - \langle \mathbf{C}^{\mathsf{T}} \rangle \cancel{\langle \mathbf{R}^{-1} \rangle \langle \mathbf{R}^{-1} \rangle^{-1}} \langle \mathbf{R}^{-1} \rangle \langle \mathbf{C} \rangle\big)^{-1} \tag{55}$$

$$= \big(\mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC} + \cancel{\langle \mathbf{C}^{\mathsf{T}} \rangle \langle \mathbf{R}^{-1} \rangle \langle \mathbf{C} \rangle} - \cancel{\langle \mathbf{C}^{\mathsf{T}} \rangle \langle \mathbf{R}^{-1} \rangle \langle \mathbf{C} \rangle}\big)^{-1} \tag{56}$$

$$= \big(\mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC}\big)^{-1}. \tag{57}$$

The Woodbury matrix identity is used to get the following preferred form for $\mathbf{F}_{11}^{-1}$,

$$\mathbf{F}_{11}^{-1} = \big(\mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC}\big)^{-1} \tag{58}$$

$$= \mathbf{P}_{n-1} - \mathbf{P}_{n-1}\big(\mathbf{I} + \boldsymbol{\Sigma}_{CRC} \mathbf{P}_{n-1}\big)^{-1} \boldsymbol{\Sigma}_{CRC} \mathbf{P}_{n-1} \tag{59}$$

$$= \big(\mathbf{I} - \mathbf{P}_{n-1}\big(\mathbf{I} + \boldsymbol{\Sigma}_{CRC} \mathbf{P}_{n-1}\big)^{-1} \boldsymbol{\Sigma}_{CRC}\big) \mathbf{P}_{n-1} \tag{60}$$

$$= \mathbf{L}_{n-1} \mathbf{P}_{n-1} \tag{61}$$

where we have defined

$$\mathbf{L}_{n-1} = \mathbf{I} - \mathbf{P}_{n-1}\big(\mathbf{I} + \boldsymbol{\Sigma}_{CRC} \mathbf{P}_{n-1}\big)^{-1} \boldsymbol{\Sigma}_{CRC}. \tag{62}$$

Substituting in and simplifying the expression for the covariance $\mathbf{S}_n$,

$$\mathbf{S}_n = \boldsymbol{\Omega}_{22}^{-1} + \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\Omega}_{21} \mathbf{F}_{11}^{-1} \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-1} \tag{63}$$

$$= \langle \mathbf{R}^{-1} \rangle^{-1} + \cancel{\langle \mathbf{R}^{-1} \rangle^{-1} \langle \mathbf{R}^{-1} \rangle} \langle \mathbf{C} \rangle \mathbf{L}_{n-1} \mathbf{P}_{n-1} \langle \mathbf{C}^{\mathsf{T}} \rangle \cancel{\langle \mathbf{R}^{-1} \rangle \langle \mathbf{R}^{-1} \rangle^{-1}} \tag{64}$$

$$= \langle \mathbf{R} \rangle + \langle \mathbf{C} \rangle \mathbf{L}_{n-1} \mathbf{P}_{n-1} \langle \mathbf{C}^{\mathsf{T}} \rangle. \tag{65}$$

Next, we derive the expression for the mean $\widehat{\mathbf{y}}_n$. Completing the square for $\mathbf{y}_n$ and $\mathbf{x}_n$, the linear terms are factored as

$$\boldsymbol{\ell}_1 = \mathbf{P}_{n-1}^{-1} \mathbf{m}_{n-1} - \langle \mathbf{C}^{\mathsf{T}} \mathbf{R}^{-1} \mathbf{D} \rangle \mathbf{u}_n \tag{66}$$

$$= \mathbf{P}_{n-1}^{-1} \mathbf{m}_{n-1} - \big(\langle \mathbf{C}^{\mathsf{T}} \rangle \langle \mathbf{R}^{-1} \rangle \langle \mathbf{D} \rangle + \boldsymbol{\Sigma}_{CRD}\big) \mathbf{u}_n \tag{67}$$

$$\boldsymbol{\ell}_2 = \langle \mathbf{R}^{-1} \rangle \langle \mathbf{D} \rangle \mathbf{u}_n. \tag{68}$$

The mean is then given by

$$\widehat{\mathbf{y}}_n = \mathbf{S}_n\big(\boldsymbol{\ell}_2 - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\ell}_1\big). \tag{69}$$

Using the Schur complement, we get

$$\begin{aligned}
\mathbf{S}_n\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} &= \mathbf{F}_{22}^{-1}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} && (70)\\
&= \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\mathbf{F}_{11}^{-1} && (71)\\
&= \cancel{\langle\mathbf{R}^{-1}\rangle^{-1}\langle\mathbf{R}^{-1}\rangle}\langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{P}_{n-1} && (72)\\
&= \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{P}_{n-1}. && (73)
\end{aligned}$$

Substituting this term back into the expression for $\widehat{\mathbf{y}}_n$ and simplifying, we get

$$\begin{aligned}
\widehat{\mathbf{y}}_n &= \mathbf{S}_n\big(\boldsymbol{\ell}_2 + \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\ell}_1\big) && (74)\\
&= \mathbf{S}_n\boldsymbol{\ell}_2 + \mathbf{S}_n\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\ell}_1 && (75)\\
&= \big(\langle\mathbf{R}\rangle + \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}^{\mathsf{T}}\rangle\big)\langle\mathbf{R}^{-1}\rangle\langle\mathbf{D}\rangle\mathbf{u}_n && (76)\\
&\quad + \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{P}_{n-1}\big(\mathbf{P}_{n-1}^{-1}\mathbf{m}_{n-1} - \big(\langle\mathbf{C}^{\mathsf{T}}\rangle\langle\mathbf{R}^{-1}\rangle\langle\mathbf{D}\rangle + \boldsymbol{\Sigma}_{CRD}\big)\mathbf{u}_n\big) && (77)\\
&= \langle\mathbf{D}\rangle\mathbf{u}_n + \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{m}_{n-1} - \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{P}_{n-1}\boldsymbol{\Sigma}_{CRD}\mathbf{u}_n. && (78)
\end{aligned}$$

In summary, the marginalization over $\mathbf{x}_n$ gives the marginal likelihood (filtered output probability)

$$q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1}) = \mathcal{N}(\mathbf{y}|\widehat{\mathbf{y}}_n, \mathbf{S}_n) \tag{79}$$

where

$$\begin{aligned}
\mathbf{L}_{n-1} &= \mathbf{I} - \mathbf{P}_{n-1}\big(\mathbf{I} + \boldsymbol{\Sigma}_{CRC}\mathbf{P}_{n-1}\big)^{-1}\boldsymbol{\Sigma}_{CRC} && (80)\\
\widehat{\mathbf{y}}_n &= \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\left(\mathbf{m}_{n-1} - \mathbf{P}_{n-1}\boldsymbol{\Sigma}_{CRD}\mathbf{u}_n\right) + \langle\mathbf{D}\rangle\mathbf{u}_n && (81)\\
\mathbf{S}_n &= \langle\mathbf{R}\rangle + \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}^{\mathsf{T}}\rangle. && (82)
\end{aligned}$$

A.3. Marginal posterior

Returning to the marginal posterior distribution, we substitute in the previously derived results and get

$$\begin{aligned}
q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n}) &= \frac{p(\mathbf{y}_n|\mathbf{x}_n)q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1})}{q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1})} && (83)\\
&= \frac{\mathcal{N}(\mathbf{y}_n|\mathbf{C}\mathbf{x}_n + \mathbf{D}\mathbf{u}_n, \mathbf{R})\mathcal{N}(\mathbf{x}_n|\mathbf{m}_{n-1}, \mathbf{P}_{n-1})}{\mathcal{N}(\mathbf{y}|\widehat{\mathbf{y}}_n, \mathbf{S}_n)} && (84)\\
&= \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_n, \mathbf{V}_n). && (85)
\end{aligned}$$

The covariance $\mathbf{V}_n$ is found by completing the square for $\mathbf{x}_n$,

$$\begin{aligned}
\mathbf{V}_n &= \big(\mathbf{P}_{n-1}^{-1} + \langle\mathbf{C}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{C}\rangle\big)^{-1} && (86)\\
&= \big(\mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC} + \langle\mathbf{C}^{\mathsf{T}}\rangle\langle\mathbf{R}^{-1}\rangle\langle\mathbf{C}\rangle\big)^{-1}. && (87)
\end{aligned}$$

The Woodbury matrix identity can again be applied to get this result into the preferred form. Let

$$\begin{aligned}
\boldsymbol{\Omega}_{11} &= -\langle\mathbf{R}\rangle && (88)\\
\boldsymbol{\Omega}_{12} &= \langle\mathbf{C}\rangle && (89)\\
\boldsymbol{\Omega}_{21} &= \langle\mathbf{C}^{\mathsf{T}}\rangle && (90)\\
\boldsymbol{\Omega}_{22} &= \mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC} && (91)
\end{aligned}$$

then we re-write the expression for $\mathbf{V}_n$ and use the identity

$$\mathbf{V}_n = \big(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\big)^{-1} \tag{92}$$
$$= \boldsymbol{\Omega}_{22}^{-1} + \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\mathbf{F}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1} \tag{93}$$

where

$$\mathbf{F}_{11}^{-1} = \big(\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\big)^{-1} \tag{94}$$
$$= \big(-\langle\mathbf{R}\rangle - \langle\mathbf{C}\rangle(\mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC})^{-1}ev\mathbf{C}^{\mathsf{T}}\big)^{-1} \tag{95}$$
$$= -\big(\langle\mathbf{R}\rangle + \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}^{\mathsf{T}}\rangle\big)^{-1} \tag{96}$$
$$= -\mathbf{S}_n^{-1}. \tag{97}$$

Substituting back into the expression for $\mathbf{V}_n$ gives

$$\mathbf{V}_n = \boldsymbol{\Omega}_{22}^{-1} + \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\mathbf{F}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1} \tag{98}$$
$$= (\mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC})^{-1} - (\mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC})^{-1}\langle\mathbf{C}^{\mathsf{T}}\rangle\mathbf{S}_n^{-1}\langle\mathbf{C}\rangle(\mathbf{P}_{n-1}^{-1} + \boldsymbol{\Sigma}_{CRC})^{-1} \tag{99}$$
$$= \mathbf{L}_{n-1}\mathbf{P}_{n-1} - \mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}^{\mathsf{T}}\rangle\mathbf{S}_n^{-1}\langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{P}_{n-1} \tag{100}$$
$$= \big(\mathbf{I} - \mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}^{\mathsf{T}}\rangle\mathbf{S}_n^{-1}\langle\mathbf{C}\rangle\big)\mathbf{L}_{n-1}\mathbf{P}_{n-1} \tag{101}$$
$$= \big(\mathbf{I} - \mathbf{K}_n\langle\mathbf{C}\rangle\big)\mathbf{L}_{n-1}\mathbf{P}_{n-1} \tag{102}$$

where we have defined

$$\mathbf{K}_n = \mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}^{\mathsf{T}}\rangle\mathbf{S}_n^{-1}. \tag{103}$$

In the non-Bayesian scenario (when the parameters are deterministic), $\mathbf{K}_n$ is called the *Kalman gain*. Lastly, the mean is found by completing the square over $\mathbf{x}_n$. Collecting the linear terms over $\mathbf{x}_n$ gives

$$\boldsymbol{\mu}_n = \mathbf{V}_n\big(\mathbf{P}_{n-1}^{-1}\mathbf{m}_{n-1} - \big(\langle\mathbf{C}^{\mathsf{T}}\rangle\langle\mathbf{R}^{-1}\rangle\langle\mathbf{D}\rangle + \boldsymbol{\Sigma}_{CRD}\big)\mathbf{u}_n + \langle\mathbf{C}^{\mathsf{T}}\mathbf{R}^{-1}\rangle\mathbf{y}_n\big). \tag{104}$$

Each term in the previous equation is addressed in order. First, we have

$$\mathbf{V}_n\mathbf{P}_{n-1}^{-1}\mathbf{m}_{n-1} = \big(\mathbf{I} - \mathbf{K}_n\langle\mathbf{C}\rangle\big)\mathbf{L}_{n-1}\cancel{\mathbf{P}_{n-1}\mathbf{P}_{n-1}^{-1}}\mathbf{m}_{n-1} \tag{105}$$
$$= \big(\mathbf{I} - \mathbf{K}_n\langle\mathbf{C}\rangle\big)\mathbf{L}_{n-1}\mathbf{m}_{n-1} \tag{106}$$
$$= \mathbf{L}_{n-1}\mathbf{m}_{n-1} - \mathbf{K}_n\langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{m}_{n-1}. \tag{107}$$

Next, $\mathbf{V}_n\langle\mathbf{C}^{\mathsf{T}}\rangle\langle\mathbf{R}^{-1}\rangle$ can be simplified. Recall from the conditional distribution that $\boldsymbol{\Omega}_{12} = \langle\mathbf{C}^{\mathsf{T}}\rangle\langle\mathbf{R}^{-1}\rangle$, and $\mathbf{V}_n = \boldsymbol{\Omega}_{11}^{-1} = \mathbf{P}_{n-1}^{-1} + \langle\mathbf{C}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{C}\rangle$. Then,

$$\mathbf{V}_n\langle\mathbf{C}^{\mathsf{T}}\rangle\langle\mathbf{R}^{-1}\rangle = \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} \tag{108}$$
$$= \mathbf{F}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\mathbf{F}_{22} \tag{109}$$
$$= \mathbf{L}\mathbf{P}_{n-1}\langle\mathbf{C}^{\mathsf{T}}\rangle\cancel{\langle\mathbf{R}^{-1}\rangle\langle\mathbf{R}^{-1}\rangle^{-1}}\mathbf{S}_n^{-1} \tag{110}$$
$$= \mathbf{L}\mathbf{P}_{n-1}\langle\mathbf{C}^{\mathsf{T}}\rangle\mathbf{S}_n^{-1} \tag{111}$$
$$= \mathbf{K}_n. \tag{112}$$

Substituting this term back into the expression for $\boldsymbol{\mu}_n$ and simplifying, we get

$$\boldsymbol{\mu}_n = \mathbf{V}_n\big(\mathbf{P}_{n-1}^{-1}\mathbf{m}_{n-1} - \big(\langle\mathbf{C}^{\mathsf{T}}\rangle\langle\mathbf{R}^{-1}\rangle\langle\mathbf{D}\rangle + \boldsymbol{\Sigma}_{CRD}\big)\mathbf{u}_n + \langle\mathbf{C}^{\mathsf{T}}\mathbf{R}^{-1}\rangle\mathbf{y}_n\big) \tag{113}$$
$$= \mathbf{V}_n\mathbf{P}_{n-1}^{-1}\mathbf{m}_{n-1} - \mathbf{V}_n\big(\langle\mathbf{C}^{\mathsf{T}}\rangle\langle\mathbf{R}^{-1}\rangle\langle\mathbf{D}\rangle + \boldsymbol{\Sigma}_{CRD}\big)\mathbf{u}_n + \mathbf{V}_n\langle\mathbf{C}^{\mathsf{T}}\mathbf{R}^{-1}\rangle\mathbf{y}_n\big) \tag{114}$$
$$= \mathbf{L}_{n-1}\mathbf{m}_{n-1} - \mathbf{K}_n\langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{m}_{n-1} - \mathbf{K}_n\langle\mathbf{D}\rangle\mathbf{u}_n - \mathbf{V}_n\boldsymbol{\Sigma}_{CRD}\mathbf{u}_n + \mathbf{K}_n\mathbf{y}_n \tag{115}$$
$$= \mathbf{L}_{n-1}\mathbf{m}_{n-1} + \mathbf{K}_n\big(\mathbf{y}_n - \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{m}_{n-1} - \langle\mathbf{D}\rangle\mathbf{u}_n\big) - \mathbf{V}_n\boldsymbol{\Sigma}_{CRD}\mathbf{u}_n \tag{116}$$
$$= \mathbf{L}_{n-1}\mathbf{m}_{n-1} + \mathbf{K}_n(\mathbf{y}_n - \widehat{\mathbf{y}}_n) - \mathbf{L}_{n-1}\mathbf{P}_{n-1}\boldsymbol{\Sigma}_{CRD}\mathbf{u}_n. \tag{117}$$

In summary, the forward pass computes the statistics of the marginal posterior probability, $\forall n \in [1..N]$:

$$q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n}) = \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_n, \mathbf{V}_n) \tag{118}$$

where

$$\mathbf{K}_n = \mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}^{\mathsf{T}}\rangle\mathbf{S}_n^{-1} \tag{119}$$

$$\boldsymbol{\mu}_n = \mathbf{L}_{n-1}\left(\mathbf{m}_{n-1} - \mathbf{P}_{n-1}\boldsymbol{\Sigma}_{CRD}\mathbf{u}_n\right) + \mathbf{K}_n(\mathbf{y}_n - \widehat{\mathbf{y}}_n) \tag{120}$$

$$\mathbf{V}_n = \big(\mathbf{I} - \mathbf{K}_n\langle\mathbf{C}\rangle\big)\mathbf{L}_{n-1}\mathbf{P}_{n-1}. \tag{121}$$

This concludes the derivation of the forward pass.

B. BACKWARD PASS (SMOOTHING) DERIVATION

The backward pass computes the sufficient statistics of the marginal [1]

$$q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{Y}) = \frac{q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n})}{q_{\mathbf{x}}(\mathbf{y}_{n+1}|\mathbf{y}_{1:n})} \int \frac{p(\mathbf{y}_{n+1}|\mathbf{x}_{n+1})p(\mathbf{x}_{n+1}|\mathbf{x}_n)q_{\mathbf{x}}(\mathbf{x}_{n+1}|\mathbf{Y})}{q_{\mathbf{x}}(\mathbf{x}_{n+1}|\mathbf{y}_{1:n+1})} d\mathbf{x}_{n+1} \tag{122}$$

$$= \frac{q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n})}{\cancel{q_{\mathbf{x}}(\mathbf{y}_{n+1}|\mathbf{y}_{1:n})}} \int \frac{\cancel{p(\mathbf{y}_{n+1}|\mathbf{x}_{n+1})}p(\mathbf{x}_{n+1}|\mathbf{x}_n)p(\mathbf{x}_{n+1}|\mathbf{Y})\cancel{q_{\mathbf{x}}(\mathbf{y}_{n+1}|\mathbf{y}_{1:n})}}{\cancel{p(\mathbf{y}_{n+1}|\mathbf{x}_{n+1})}\int p(\mathbf{x}_{n+1}|\mathbf{x}_n)q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n})d\mathbf{x}_n} d\mathbf{x}_{n+1} \tag{123}$$

$$= q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n}) \int \frac{p(\mathbf{x}_{n+1}|\mathbf{x}_n)p(\mathbf{x}_{n+1}|\mathbf{Y})}{q_{\mathbf{x}}(\mathbf{x}_{n+1}|\mathbf{y}_{1:n})} d\mathbf{x}_{n+1} \tag{124}$$

$$= \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_n, \mathbf{V}_n) \int \frac{\mathcal{N}(\mathbf{x}_{n+1}|\mathbf{A}\mathbf{x}_n + \mathbf{B}\mathbf{u}_{n+1}, \mathbf{Q})\mathcal{N}(\mathbf{x}_{n+1}|\widehat{\boldsymbol{\mu}}_{n+1}, \widehat{\mathbf{V}}_{n+1})}{\mathcal{N}(\mathbf{x}_{n+1}|\mathbf{m}_n, \mathbf{P}_n)} d\mathbf{x}_{n+1} \tag{125}$$

$$= \mathcal{N}(\mathbf{x}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n). \tag{126}$$

The square is completed over $\mathbf{x}_n$ and $\mathbf{x}_{n-1}$ to factor the quadratic terms

$$\boldsymbol{\Omega}_{22} = \mathbf{V}_n^{-1} + \langle \mathbf{A}^\mathsf{T}\mathbf{Q}^{-1}\mathbf{A}\rangle \tag{127}$$

$$= \mathbf{V}_n^{-1} + \boldsymbol{\Sigma}_{AQA} + \langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle \tag{128}$$

$$\boldsymbol{\Omega}_{21} = \langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle \tag{129}$$

$$\boldsymbol{\Omega}_{12} = \langle \mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle \tag{130}$$

$$\boldsymbol{\Omega}_{11} = \widehat{\mathbf{V}}_{n+1}^{-1} + \langle \mathbf{Q}^{-1}\rangle - \mathbf{P}_n^{-1} \tag{131}$$

$$= \widehat{\mathbf{V}}_{n+1}^{-1} + \cancel{\langle \mathbf{Q}^{-1}\rangle} - \big(\cancel{\langle \mathbf{Q}^{-1}\rangle} - \langle \mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle\big(\mathbf{V}_n^{-1} + \langle \mathbf{A}^\mathsf{T}\mathbf{Q}^{-1}\mathbf{A}\rangle\big)^{-1}\langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle\big) \tag{132}$$

$$= \widehat{\mathbf{V}}_{n+1}^{-1} + \langle \mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle\Big(\mathbf{V}_n^{-1} + \boldsymbol{\Sigma}_{AQA} + \langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle\Big)^{-1}\langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle \tag{133}$$

and the linear terms

$$\boldsymbol{\ell}_2 = \mathbf{V}_n^{-1}\boldsymbol{\mu}_n - \langle \mathbf{A}^\mathsf{T}\mathbf{Q}^{-1}\mathbf{B}\rangle\mathbf{u}_{n+1} \tag{134}$$

$$= \mathbf{V}_n^{-1}\boldsymbol{\mu}_n - \big(\langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{B}\rangle + \boldsymbol{\Sigma}_{AQB}\big)\mathbf{u}_{n+1} \tag{135}$$

$$\boldsymbol{\ell}_1 = \widehat{\mathbf{V}}_{n+1}^{-1}\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{P}_n^{-1}\mathbf{m}_n + \langle \mathbf{Q}^{-1}\mathbf{B}\rangle\mathbf{u}_{n+1}. \tag{136}$$

The covariance and mean are

$$\widehat{\mathbf{V}}_n = \big(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\big)^{-1} \tag{137}$$

$$\widehat{\boldsymbol{\mu}}_n = \widehat{\mathbf{V}}_n\big(\boldsymbol{\ell}_2 - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\ell}_1\big). \tag{138}$$

We will use results from matrix theory to get these results into a preferred form, reminiscent of the Rauch-Tung-Striebel (RTS) smoother [2]. First the covariance. Using the Schur complement, we get

$$\widehat{\mathbf{V}}_n = \big(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\big)^{-1} \tag{139}$$

$$= \boldsymbol{\Omega}_{22}^{-1} + \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\mathbf{F}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}. \tag{140}$$

The term $\mathbf{F}_{11}^{-1}$ is reduced to

$$\mathbf{F}_{11}^{-1} = \big(\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\big)^{-1} \tag{141}$$

$$= \Bigg(\widehat{\mathbf{V}}_{n+1}^{-1} + \langle \mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle\Big(\mathbf{V}_n^{-1} + \boldsymbol{\Sigma}_{AQA} + \cancel{\langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle}\Big)^{-1}\langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle \tag{142}$$

$$\cancel{- \langle \mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle\Big(\mathbf{V}_n^{-1} + \boldsymbol{\Sigma}_{AQA} + \langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{A}\rangle\Big)^{-1}\langle \mathbf{A}^\mathsf{T}\rangle\langle\mathbf{Q}^{-1}\rangle}\Bigg)^{-1} \tag{143}$$

$$= \big(\widehat{\mathbf{V}}_{n+1}^{-1}\big)^{-1} \tag{144}$$

$$= \widehat{\mathbf{V}}_{n+1}. \tag{145}$$

Next, the inverse of $\boldsymbol{\Omega}_{22}$ is analytically expressed using the Woodbury matrix identity. Defining

$$\boldsymbol{\Lambda}_{11} = -\langle \mathbf{Q} \rangle \tag{146}$$
$$\boldsymbol{\Lambda}_{12} = \langle \mathbf{A} \rangle \tag{147}$$
$$\boldsymbol{\Lambda}_{21} = \langle \mathbf{A}^\mathsf{T} \rangle \tag{148}$$
$$\boldsymbol{\Lambda}_{22} = \mathbf{V}_n^{-1} + \boldsymbol{\Sigma}_{AQA} \tag{149}$$

the inverse is then

$$\boldsymbol{\Omega}_{22}^{-1} = \big(\boldsymbol{\Lambda}_{22} - \boldsymbol{\Lambda}_{21}\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}\big)^{-1} \tag{150}$$
$$= \boldsymbol{\Lambda}_{22}^{-1} + \boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}\mathbf{O}_{11}^{-1}\boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}. \tag{151}$$

The term $\boldsymbol{\Lambda}_{22}^{-1}$ is simplified using equation (20) from the forward pass:

$$\boldsymbol{\Lambda}_{22}^{-1} = \big(\mathbf{V}_n^{-1} + \boldsymbol{\Sigma}_{AQA}\big)^{-1} \tag{152}$$
$$= \mathbf{G}_n\mathbf{V}_n. \tag{153}$$

Likewise, the term $\mathbf{O}_{11}^{-1}$ is given by

$$\mathbf{O}_{11}^{-1} = \big(\boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}\big)^{-1} \tag{154}$$
$$= \big(-\langle \mathbf{Q} \rangle - \langle \mathbf{A} \rangle \mathbf{G}_n\mathbf{V}_n\langle \mathbf{A}^\mathsf{T} \rangle\big)^{-1} \tag{155}$$
$$= -\mathbf{P}_n^{-1}. \tag{156}$$

These results are substituted into the equation for $\boldsymbol{\Omega}_{22}^{-1}$:

$$\boldsymbol{\Omega}_{22}^{-1} = \boldsymbol{\Lambda}_{22}^{-1} + \boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}\mathbf{O}_{11}^{-1}\boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1} \tag{157}$$
$$= \mathbf{G}_n\mathbf{V}_n - \mathbf{G}_n\mathbf{V}_n\langle \mathbf{A}^\mathsf{T} \rangle \mathbf{P}_n^{-1}\langle \mathbf{A} \rangle \mathbf{G}_n\mathbf{V}_n. \tag{158}$$

Lastly, we can use another matrix identity to simplify the expression $\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}$.

$$\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21} = \Big(\mathbf{V}_n^{-1} + \boldsymbol{\Sigma}_{AQA} + \langle \mathbf{A}^\mathsf{T} \rangle \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{A} \rangle\Big)^{-1} \langle \mathbf{A}^\mathsf{T} \rangle \langle \mathbf{Q}^{-1} \rangle \tag{159}$$
$$= \big(\mathbf{V}_n^{-1} + \boldsymbol{\Sigma}_{AQA}\big)^{-1} \langle \mathbf{A}^\mathsf{T} \rangle \Big(\langle \mathbf{Q} \rangle + \langle \mathbf{A} \rangle \big(\mathbf{V}_n^{-1} + \boldsymbol{\Sigma}_{AQA}\big)^{-1} \langle \mathbf{A}^\mathsf{T} \rangle\Big)^{-1} \tag{160}$$
$$= \mathbf{G}_n\mathbf{V}_n \langle \mathbf{A}^\mathsf{T} \rangle \Big(\langle \mathbf{Q} \rangle + \langle \mathbf{A} \rangle \mathbf{G}_n\mathbf{V}_n \langle \mathbf{A}^\mathsf{T} \rangle\Big)^{-1} \tag{161}$$
$$= \mathbf{G}_n\mathbf{V}_n \langle \mathbf{A}^\mathsf{T} \rangle \mathbf{P}_n^{-1} \tag{162}$$

Putting this altogether, and noting that $\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1} = \big(\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\big)^\mathsf{T}$, the covariance from the backward pass at time n is given by

$$\widehat{\mathbf{V}}_n = \boldsymbol{\Omega}_{22}^{-1} + \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\mathbf{F}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1} \tag{163}$$
$$= \mathbf{G}_n\mathbf{V}_n - \mathbf{G}_n\mathbf{V}_n\langle \mathbf{A}^\mathsf{T} \rangle \mathbf{P}_n^{-1}\langle \mathbf{A} \rangle \mathbf{G}_n\mathbf{V}_n + \mathbf{G}_n\mathbf{V}_n\langle \mathbf{A}^\mathsf{T} \rangle \mathbf{P}_n^{-1}\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle \mathbf{A} \rangle \mathbf{G}_n\mathbf{V}_n \tag{164}$$
$$= \mathbf{G}_n\mathbf{V}_n + \mathbf{G}_n\mathbf{V}_n\langle \mathbf{A}^\mathsf{T} \rangle \mathbf{P}_n^{-1}\big(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n\big)\mathbf{P}_n^{-1}\langle \mathbf{A} \rangle \mathbf{G}_n\mathbf{V}_n \tag{165}$$
$$= \mathbf{G}_n\mathbf{V}_n + \mathbf{J}_n\big(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n\big)\mathbf{J}_n^\mathsf{T} \tag{166}$$

where we have defined

$$\mathbf{J}_n = \mathbf{G}_n\mathbf{V}_n\langle \mathbf{A}^\mathsf{T} \rangle \mathbf{P}_n^{-1}. \tag{167}$$

Lastly, the mean $\widehat{\boldsymbol{\mu}}_n$ is given by

$$\widehat{\boldsymbol{\mu}}_n = \widehat{\mathbf{V}}_n\big(\boldsymbol{\ell}_2 + \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\ell}_1\big) \tag{168}$$
$$= \widehat{\mathbf{V}}_n\mathbf{V}_n^{-1}\boldsymbol{\mu}_n - \widehat{\mathbf{V}}_n\big(\langle \mathbf{A}^\mathsf{T} \rangle \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{B} \rangle + \boldsymbol{\Sigma}_{AQB}\big)\mathbf{u}_{n+1} + \widehat{\mathbf{V}}_n\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\ell}_1 \tag{169}$$
$$= \widehat{\mathbf{V}}_n\mathbf{V}_n^{-1}\boldsymbol{\mu}_n - \widehat{\mathbf{V}}_n\langle \mathbf{A}^\mathsf{T} \rangle \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{B} \rangle \mathbf{u}_{n+1} - \widehat{\mathbf{V}}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} + \widehat{\mathbf{V}}_n\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\ell}_1. \tag{170}$$

The left-most term, $\widehat{\mathbf{V}}_n\mathbf{V}_n^{-1}\boldsymbol{\mu}_n$, is

$$\widehat{\mathbf{V}}_n\mathbf{V}_n^{-1}\boldsymbol{\mu}_n = \Big(\mathbf{G}_n\mathbf{V}_n + \mathbf{J}_n\big(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n\big)\mathbf{J}_n^{\mathsf{T}}\Big)\mathbf{V}_n^{-1}\boldsymbol{\mu}_n \tag{171}$$

$$= \Big(\mathbf{G}_n\cancel{\mathbf{V}_n} + \mathbf{J}_n\big(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n\big)\mathbf{P}_n^{-1}\langle\mathbf{A}\rangle\mathbf{G}_n\cancel{\mathbf{V}_n}\Big)\cancel{\mathbf{V}_n^{-1}}\boldsymbol{\mu}_n \tag{172}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\big(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n\big)\mathbf{P}_n^{-1}\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n \tag{173}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n - \mathbf{J}_n\cancel{\mathbf{P}_n}\cancel{\mathbf{P}_n^{-1}}\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n \tag{174}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n - \mathbf{J}_n\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n. \tag{175}$$

Next, $\widehat{\mathbf{V}}_n\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}$ can be evaluated using the Schur complement:

$$\widehat{\mathbf{V}}_n\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} = \mathbf{F}_{22}^{-1}\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} = \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\mathbf{F}_{11}^{-1} \tag{176}$$

$$= \mathbf{G}_n\mathbf{V}_n\langle\mathbf{A}^{\mathsf{T}}\rangle\mathbf{P}_n^{-1}\widehat{\mathbf{V}}_{n+1} \tag{177}$$

$$= \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}. \tag{178}$$

Likewise, we can use the Schur complement to evaluate $\widehat{\mathbf{V}}_n\langle\mathbf{A}^{\mathsf{T}}\rangle\langle\mathbf{Q}^{-1}\rangle$.

$$\widehat{\mathbf{V}}_n\langle\mathbf{A}^{\mathsf{T}}\rangle\langle\mathbf{Q}^{-1}\rangle = \mathbf{F}_{22}^{-1}\boldsymbol{\Omega}_{21} = \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}\mathbf{F}_{11}^{-1}\boldsymbol{\Omega}_{11} \tag{179}$$

$$= \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\boldsymbol{\Omega}_{11} \tag{180}$$

We expand the last term $\widehat{\mathbf{V}}_n\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\ell}_1$ as follows.

$$\widehat{\mathbf{V}}_n\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\ell}_1 = \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\big(\widehat{\mathbf{V}}_{n+1}^{-1}\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{P}_n^{-1}\mathbf{m}_n + \langle\mathbf{Q}^{-1}\mathbf{B}\rangle\mathbf{u}_{n+1}\big) \tag{181}$$

$$= \mathbf{J}_n\cancel{\widehat{\mathbf{V}}_{n+1}}\cancel{\widehat{\mathbf{V}}_{n+1}^{-1}}\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\mathbf{m}_n + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\langle\mathbf{Q}^{-1}\mathbf{B}\rangle\mathbf{u}_{n+1} \tag{182}$$

$$= \mathbf{J}_n\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\mathbf{m}_n + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\langle\mathbf{Q}^{-1}\mathbf{B}\rangle\mathbf{u}_{n+1} \tag{183}$$

The second term in the previous equation $\mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\mathbf{m}_n$ is further expanded to give

$$\mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\mathbf{m}_n = \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}(\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n + \big(\langle\mathbf{B}\rangle - \langle\mathbf{A}\rangle\mathbf{G}_n\mathbf{V}_n\boldsymbol{\Sigma}_{AQB}\big)\mathbf{u}_{n+1}) \tag{184}$$

$$= \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle\mathbf{B}\rangle\mathbf{u}_{n+1} - \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle\mathbf{A}\rangle\mathbf{G}_n\mathbf{V}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} \tag{185}$$

$$= \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle\mathbf{B}\rangle\mathbf{u}_{n+1} - \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}}\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1}. \tag{186}$$

Now we substitute all these terms back into equation (170) and get

$$\widehat{\boldsymbol{\mu}}_n = \widehat{\mathbf{V}}_n\mathbf{V}_n^{-1}\boldsymbol{\mu}_n - \widehat{\mathbf{V}}_n\langle\mathbf{A}^{\mathsf{T}}\rangle\langle\mathbf{Q}^{-1}\rangle\langle\mathbf{B}\rangle\mathbf{u}_{n+1} - \widehat{\mathbf{V}}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} + \widehat{\mathbf{V}}_n\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\ell}_1 \tag{187}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \cancel{\mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n} - \mathbf{J}_n\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n - \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\boldsymbol{\Omega}_{11}\langle\mathbf{B}\rangle\mathbf{u}_{n+1} - \widehat{\mathbf{V}}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} + \mathbf{J}_n\widehat{\boldsymbol{\mu}}_{n+1} \tag{188}$$

$$- \cancel{\mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n} - \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{P}_n^{-1}\langle\mathbf{B}\rangle\mathbf{u}_{n+1} + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}}\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\langle\mathbf{Q}^{-1}\mathbf{B}\rangle\mathbf{u}_{n+1} \tag{189}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\Big(\widehat{\boldsymbol{\mu}}_{n+1} - \langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n\Big) - \widehat{\mathbf{V}}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}}\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} \tag{190}$$

$$+ \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\Big(\langle\mathbf{Q}^{-1}\rangle - \mathbf{P}_n^{-1} - \boldsymbol{\Omega}_{11}\Big)\langle\mathbf{B}\rangle\mathbf{u}_{n+1} \tag{191}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\Big(\widehat{\boldsymbol{\mu}}_{n+1} - \langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n\Big) - \widehat{\mathbf{V}}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}}\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} \tag{192}$$

$$+ \mathbf{J}_n\cancel{\widehat{\mathbf{V}}_{n+1}}\Big(\cancel{-\widehat{\mathbf{V}}_{n+1}^{-1}}\Big)\langle\mathbf{B}\rangle\mathbf{u}_{n+1} \tag{193}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\Big(\widehat{\boldsymbol{\mu}}_{n+1} - \langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n\Big) - \widehat{\mathbf{V}}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}}\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} - \mathbf{J}_n\langle\mathbf{B}\rangle\mathbf{u}_{n+1} \tag{194}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\big(\widehat{\boldsymbol{\mu}}_{n+1} - \langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n - \langle\mathbf{B}\rangle\mathbf{u}_{n+1} + \widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}}\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1}\big) - \widehat{\mathbf{V}}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1}. \tag{195}$$

Further simplification and rearrangement of terms gives

$$\widehat{\boldsymbol{\mu}}_n = \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\big(\widehat{\boldsymbol{\mu}}_{n+1} - \langle\mathbf{A}\rangle\mathbf{G}_n\boldsymbol{\mu}_n - \langle\mathbf{B}\rangle\mathbf{u}_{n+1} + \widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}}\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1}\big) - \widehat{\mathbf{V}}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} \tag{196}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\big(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{m}_n - \langle\mathbf{A}\rangle\mathbf{G}_n\mathbf{V}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} + \widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}}\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1}\big) - \widehat{\mathbf{V}}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} \tag{197}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\big(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{m}_n\big) - \mathbf{J}_n\langle\mathbf{A}\rangle\mathbf{G}_n\mathbf{V}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} + \mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}}\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} - \widehat{\mathbf{V}}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} \tag{198}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\big(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{m}_n\big) + \Big(\mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}} - \mathbf{J}_n\langle\mathbf{A}\rangle\mathbf{G}_n\mathbf{V}_n - \widehat{\mathbf{V}}_n\Big)\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} \tag{199}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\big(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{m}_n\big) + \Big(\mathbf{J}_n\widehat{\mathbf{V}}_{n+1}\mathbf{J}_n^{\mathsf{T}} - \mathbf{J}_n\mathbf{P}_n\mathbf{J}_n^{\mathsf{T}} - \widehat{\mathbf{V}}_n\Big)\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} \tag{200}$$

$$= \mathbf{G}_n\boldsymbol{\mu}_n + \mathbf{J}_n\big(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{m}_n\big) - \mathbf{G}_n\mathbf{V}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1} \tag{201}$$

$$= \mathbf{G}_n\big(\boldsymbol{\mu}_n - \mathbf{V}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1}\big) + \mathbf{J}_n\big(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{m}_n\big). \tag{202}$$

In summary, the backward pass computes the statistics of the smoothed marginal posterior probability:

$$q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{Y}) = \mathcal{N}(\mathbf{x}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n) \tag{203}$$

where

$$\mathbf{J}_n = \mathbf{G}_n\mathbf{V}_n\langle\mathbf{A}^{\mathsf{T}}\rangle\mathbf{P}_n^{-1} \tag{204}$$

$$\widehat{\boldsymbol{\mu}}_n = \mathbf{G}_n\big(\boldsymbol{\mu}_n - \mathbf{V}_n\boldsymbol{\Sigma}_{AQB}\mathbf{u}_{n+1}\big) + \mathbf{J}_n\big(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{m}_n\big) \tag{205}$$

$$\widehat{\mathbf{V}}_n = \mathbf{G}_n\mathbf{V}_n + \mathbf{J}_n\big(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n\big)\mathbf{J}_n^{\mathsf{T}}. \tag{206}$$

This concludes the derivation of the backward pass (smoother) for the Bayesian linear dynamical system.

C. APPLICATION: VARIATIONAL M STEP DERIVATION

C.1. Joint distribution

The parameters of the Bayesian frequency estimation model are $\boldsymbol{\theta} = (\boldsymbol{\nu}, \boldsymbol{\tau}, \rho)$. The joint distribution over the variables and parameters is

$$p(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) = p(\mathbf{Y}|\mathbf{X}, \rho)p(\mathbf{X}|\boldsymbol{\nu}, \boldsymbol{\tau})p(\boldsymbol{\nu}|\boldsymbol{\tau})p(\boldsymbol{\tau})p(\rho) \tag{207}$$

where the conditional and marginal distributions are

$$p(\mathbf{Y}|\mathbf{X}, \rho) = \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{C}\mathbf{x}_n, R) \tag{208}$$

$$p(\mathbf{X}|\boldsymbol{\nu}, \boldsymbol{\tau}) = \mathcal{N}(\mathbf{x}_1|\mathbf{m}_0, \mathbf{P}_0)\prod_{n=2}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{A}\mathbf{x}_{n-1}, \mathbf{Q}) \tag{209}$$

$$p(\boldsymbol{\nu}|\boldsymbol{\tau}) = \prod_{k=1}^{K} \mathcal{N}(\nu_k|0, \alpha_k^{-1}\tau_k^{-1}) \tag{210}$$

$$p(\boldsymbol{\tau}) = \prod_{k=1}^{K} \text{Gam}(\tau_k|e_0, i_0) \tag{211}$$

$$p(\rho) = \text{Gam}\big(\rho|r_0, s_0\big) \tag{212}$$

C.2. Optimal distribution over parameters

The log optimal distribution over the parameters factors naturally as

$$\ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \langle \ln p(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \rangle_{q_{\mathbf{x}}(\mathbf{x})} + \text{const.} \tag{213}$$

$$= \ln q_{\boldsymbol{\theta}}(\boldsymbol{\nu}, \boldsymbol{\tau}) + \ln q_{\boldsymbol{\theta}}(\rho) \tag{214}$$

The log optimal factors are derived in turn. First, we have

$$\ln q_{\boldsymbol{\theta}}(\boldsymbol{\nu}, \boldsymbol{\tau}) = \langle \ln p(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \rangle_{q_{\mathbf{x}}(\mathbf{x})} + \text{const.} \tag{215}$$

$$= \langle \ln p(\mathbf{X}|\boldsymbol{\nu}, \boldsymbol{\tau}) \rangle_{q_{\mathbf{x}}(\mathbf{x})} + \ln p(\boldsymbol{\nu}|\boldsymbol{\tau}) + \ln p(\boldsymbol{\tau}) + \text{const.} \tag{216}$$

The expected log likelihood of the latent state sequence is given by

$$\begin{aligned}\langle \ln p(\mathbf{X}|\boldsymbol{\nu}, \boldsymbol{\tau}) \rangle_{q_{\mathbf{x}}(\mathbf{x})} = &\sum_{k=1}^{K} (N-1)\ln \tau_k - \frac{\tau_k}{2}\text{Tr}\left(\sum_{n=2}^{N}\langle \mathbf{x}_{n,k}\mathbf{x}_{n,k}^{\mathsf{T}}\rangle\right) + \tau_k \text{Tr}\left(\mathbf{F}\sum_{n=2}^{N}\langle \mathbf{x}_{n-1,k}\mathbf{x}_{n,k}^{\mathsf{T}}\rangle\right) \\ &+ \tau_k\nu_k\text{Tr}\left(\mathbf{E}\sum_{n=2}^{N}\langle \mathbf{x}_{n-1,k}\mathbf{x}_{n,k}^{\mathsf{T}}\rangle\right) - \frac{\tau_k}{2}\text{Tr}\left(\mathbf{F}\sum_{n=2}^{N}\langle \mathbf{x}_{n-1,k}\mathbf{x}_{n-1,k}\rangle\mathbf{F}^{\mathsf{T}}\right) \\ &- \frac{\tau_k\nu_k^2}{2}\text{Tr}\left(\mathbf{E}\sum_{n=2}^{N}\langle \mathbf{x}_{n-1,k}\mathbf{x}_{n-1,k}\rangle\mathbf{E}^{\mathsf{T}}\right) + \tau_k\nu_k\text{Tr}\left(\mathbf{E}\sum_{n=2}^{N}\langle \mathbf{x}_{n-1,k}\mathbf{x}_{n-1,k}\rangle\mathbf{F}^{\mathsf{T}}\right) + \text{const.}\end{aligned} \tag{217}$$

The log prior distributions over $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$ are, respectively,

$$\ln p(\boldsymbol{\nu}|\boldsymbol{\tau}) = \sum_{k=1}^{K} \frac{1}{2}\ln \tau_k - \frac{\tau_k}{2}\alpha_k\nu_k^2 + \text{const.} \tag{218}$$

$$\ln p(\boldsymbol{\tau}) = \sum_{k=1}^{K} (e_0 - 1)\ln \tau_k - i_0\tau_k + \text{const.} \tag{219}$$

The optimal distribution over these parameters is of the same form as the prior distribution (Normal-Gamma):

$$q_{\boldsymbol{\theta}}(\boldsymbol{\nu}, \boldsymbol{\tau}) = q_{\boldsymbol{\nu}}(\boldsymbol{\nu}|\boldsymbol{\tau})q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=1}^{K} \mathcal{N}(\nu_k|\widehat{\nu}_k, \sigma_k\tau_k^{-1})\text{Gam}(\tau_k|\widehat{e}_k, \widehat{i}_k). \tag{220}$$

The statistics of the optimal distribution are found by completing the square for ν_k and collecting the $\ln \tau_k$ and τ_k terms from the log optimal distribution above. They are given by

$$\sigma_k = \left(\mathrm{Tr} \left(\mathbf{E} \sum_{n=2}^{N} \langle \mathbf{x}_{n-1k} \mathbf{x}_{n-1k}^{\mathsf{T}} \rangle \mathbf{E}^{\mathsf{T}} \right) + \alpha_k \right)^{-1} \tag{221}$$

$$\widehat{\nu}_k = \sigma_k \mathrm{Tr} \left(\mathbf{E} \left(\sum_{n=2}^{N} \langle \mathbf{x}_{n-1k} \mathbf{x}_{nk}^{\mathsf{T}} \rangle - \sum_{n=2}^{N} \langle \mathbf{x}_{n-1k} \mathbf{x}_{n-1k}^{\mathsf{T}} \rangle \mathbf{F}^{\mathsf{T}} \right) \right) \tag{222}$$

$$\widehat{e}_k = e_0 + N - 1 \tag{223}$$

$$\widehat{i}_k = i_0 - \frac{1}{2} \nu_k^2 \sigma_k^{-1} + \frac{1}{2} \mathrm{Tr} \Big(\mathbf{F} \sum_{n=2}^{N} \langle \mathbf{x}_{n-1k} \mathbf{x}_{n-1k}^{\mathsf{T}} \rangle \mathbf{F}^{\mathsf{T}} - 2 \mathbf{F} \sum_{n=2}^{N} \langle \mathbf{x}_{n-1k} \mathbf{x}_{nk}^{\mathsf{T}} \rangle + \sum_{n=2}^{N} \langle \mathbf{x}_{nk} \mathbf{x}_{nk}^{\mathsf{T}} \rangle \Big) \tag{224}$$

where $\mathbf{x}_{nk}$ is the kth 2×1 sub-vector of $\mathbf{x}_n$.

Next, the optimal distribution over the observation noise precision ρ is given by

$$\ln q_{\boldsymbol{\theta}}(\rho) = \langle \ln p(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \rangle_{q_{\mathbf{x}}(\mathbf{x})} + \text{const.} \tag{225}$$

$$= \langle \ln p(\mathbf{Y} | \mathbf{X}, \rho) \rangle_{q_{\mathbf{x}}(\mathbf{x})} + \ln p(\rho) + \text{const.}. \tag{226}$$

The expected log likelihood of the observed sequence is given by

$$\langle \ln p(\mathbf{Y} | \mathbf{X}, \boldsymbol{\theta}) \rangle_{q_{\mathbf{x}}(\mathbf{x})} = \frac{N}{2} \ln \rho - \frac{\rho}{2} \left(\sum_{n=1}^{N} \mathbf{y}_n \mathbf{y}_n^{\mathsf{T}} - 2 \mathbf{C} \sum_{n=1}^{N} \langle \mathbf{x}_n \rangle \mathbf{y}_n^{\mathsf{T}} + \mathbf{C} \sum_{n=1}^{N} \langle \mathbf{x}_n \mathbf{x}_n^{\mathsf{T}} \rangle \mathbf{C}^{\mathsf{T}} \right) + \text{const.} \tag{227}$$

The log prior distribution over ρ is

$$\ln p(\rho) = (r_0 - 1) \ln \rho - s_0 \rho. \tag{228}$$

As expected, ρ is Gamma-distributed:

$$q_\rho(\rho) = \mathrm{Gam}(\rho | \widehat{r}, \widehat{s}). \tag{229}$$

The statistics of the optimal distribution are found by collecting terms involving $\ln \rho$ and ρ that appear in $\ln q_{\boldsymbol{\theta}}(\rho)$.

$$\widehat{r} = r_0 + N/2 \tag{230}$$

$$\widehat{s} = s_0 + \frac{1}{2} \left(\sum_{n=1}^{N} \mathbf{y}_n \mathbf{y}_n^{\mathsf{T}} - 2 \mathbf{C} \sum_{n=1}^{N} \langle \mathbf{x}_n \rangle \mathbf{y}_n^{\mathsf{T}} + \mathbf{C} \sum_{n=1}^{N} \langle \mathbf{x}_n \mathbf{x}_n^{\mathsf{T}} \rangle \mathbf{C}^{\mathsf{T}} \right). \tag{231}$$

This concludes the variational M step.

D. QUADRATIC FORMS OF RANDOM VARIABLES

This section derives the decomposition of the quadratic form of parameters under the optimal distribution $q_{\boldsymbol{\theta}}$, which appears in the equations of the proposed filter and smoother. A comprehensive reference on quadratic forms of random variables is [3].

D.1. General result

The general decomposition is given by

$$\langle \boldsymbol{\Psi}^{\mathsf{T}} \boldsymbol{\Lambda}^{-1} \boldsymbol{\Omega} \rangle = \langle \boldsymbol{\Psi} \rangle^{\mathsf{T}} \langle \boldsymbol{\Lambda}^{-1} \rangle \langle \boldsymbol{\Omega} \rangle + \boldsymbol{\Sigma}_{\Psi\Lambda\Omega} \tag{232}$$

where we define the covariance term as

$$\boldsymbol{\Sigma}_{\Psi\Lambda\Omega} = \textstyle\sum_i \sum_j \langle \boldsymbol{\Lambda}^{-1} \rangle_{(i,j)} \text{cov}[\boldsymbol{\psi}_{(i)}, \boldsymbol{\omega}_{(j)}]. \tag{233}$$

where $\boldsymbol{\psi}_{(i)}$ denotes the transposed ith row of matrix $\boldsymbol{\Psi}$, and the expectations are taken with respect to the optimal distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. In the following subsections, we show that this general expression holds for an arbitrary definition of $p(\boldsymbol{\theta})$. First, we consider the case when the prior over the noise covariance matrix $\boldsymbol{\Lambda}$ is conjugate to the priors over the matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}$. Second, we show that it holds for non-conjugate priors as well, because the variational approximation must involve an induced factorization between non-conjugate parameters to retain analytic tractability.

D.2. Case 1: Conjugate prior structure

In the first case the prior $p(\boldsymbol{\theta})$ is constructed in such a way that the dynamics matrices are conjugate to the noise covariance. This is a common way of defining the parameter model as it simplifies the approximation, reducing the number of induced factorizations needed to keep the variational method tractable. Existing literature defines the priors in this way [4] [5] [6].

The prior over the parameters is $p(\boldsymbol{\theta}) = p(\mathbf{A}, \mathbf{B}, \mathbf{Q}) p(\mathbf{C}, \mathbf{D}, \mathbf{R})$ where

$$p(\mathbf{A}, \mathbf{B}, \mathbf{Q}) = p(\mathbf{A}|\mathbf{Q}) p(\mathbf{B}|\mathbf{Q}) p(\mathbf{Q}) \tag{234}$$

$$p(\mathbf{C}, \mathbf{D}, \mathbf{R}) = p(\mathbf{C}|\mathbf{R}) p(\mathbf{D}|\mathbf{R}) p(\mathbf{R}) \tag{235}$$

and the conditional and marginal distributions are

$$p(\mathbf{A}|\mathbf{Q}) = \prod_{h=1}^{H} \mathcal{N}(\mathbf{a}_{(h)}|\mathbf{0}, \text{diag}(\boldsymbol{\alpha}_h \tau_h)^{-1}) \tag{236}$$

$$p(\mathbf{B}|\mathbf{Q}) = \prod_{h=1}^{H} \mathcal{N}(\mathbf{b}_{(h)}|\mathbf{0}, \text{diag}(\boldsymbol{\beta}_h \tau_h)^{-1}) \tag{237}$$

$$p(\mathbf{Q}) = \prod_{h=1}^{H} \text{Gam}(\tau_h|e_0, g_0) \tag{238}$$

$$p(\mathbf{C}|\mathbf{R}) = \prod_{v=1}^{V} \mathcal{N}(\mathbf{c}_{(v)}|\mathbf{0}, \text{diag}(\boldsymbol{\gamma}_v \rho_v)^{-1}) \tag{239}$$

$$p(\mathbf{D}|\mathbf{R}) = \prod_{v=1}^{V} \mathcal{N}(\mathbf{d}_{(v)}|\mathbf{0}, \text{diag}(\boldsymbol{\delta}_v \rho_v)^{-1}) \tag{240}$$

$$p(\mathbf{R}) = \prod_{v=1}^{V} \text{Gam}\big(\rho_v|r_0, s_0\big). \tag{241}$$

The optimal distribution over the parameters that maximizes the lower bound, $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, factors naturally into

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = q_{\boldsymbol{\theta}}(\mathbf{A}, \mathbf{B}, \mathbf{Q}) q_{\boldsymbol{\theta}}(\mathbf{C}, \mathbf{D}, \mathbf{R}) \tag{242}$$

$$= q_{\boldsymbol{\theta}}(\mathbf{A}, \mathbf{B}|\mathbf{Q}) q_{\boldsymbol{\theta}}(\mathbf{Q}) q_{\boldsymbol{\theta}}(\mathbf{C}, \mathbf{D}|\mathbf{R}) q_{\boldsymbol{\theta}}(\mathbf{R}) \tag{243}$$

where

$$q_{\boldsymbol{\theta}}(\mathbf{A}, \mathbf{B}|\mathbf{Q}) = \prod_{h=1}^{H} \mathcal{N}\left(\begin{pmatrix} \mathbf{a}_{(h)} \\ \mathbf{b}_{(h)} \end{pmatrix} | \begin{pmatrix} \widehat{\mathbf{a}}_{(h)} \\ \widehat{\mathbf{b}}_{(h)} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}^h_{aa} & \boldsymbol{\Sigma}^h_{ab} \\ \boldsymbol{\Sigma}^h_{ba} & \boldsymbol{\Sigma}^h_{bb} \end{pmatrix} \tau_h^{-1} \right) \tag{244}$$

$$q_{\boldsymbol{\theta}}(\mathbf{Q}) = \prod_{h=1}^{H} \text{Gam}\left(\tau_h | \widehat{e}_h, \widehat{g}_h\right) \tag{245}$$

$$q_{\boldsymbol{\theta}}(\mathbf{C}, \mathbf{D}|\mathbf{R}) = \prod_{v=1}^{V} \mathcal{N}\left(\begin{pmatrix} \mathbf{c}_{(v)} \\ \mathbf{d}_{(v)} \end{pmatrix} | \begin{pmatrix} \widehat{\mathbf{c}}_{(v)} \\ \widehat{\mathbf{d}}_{(v)} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}^v_{cc} & \boldsymbol{\Sigma}^v_{cd} \\ \boldsymbol{\Sigma}^v_{dc} & \boldsymbol{\Sigma}^v_{dd} \end{pmatrix} \rho_v^{-1} \right) \tag{246}$$

$$q_{\boldsymbol{\theta}}(\mathbf{R}) = \prod_{v=1}^{V} \text{Gam}\left(\rho_v | \widehat{r}_v, \widehat{s}_v\right). \tag{247}$$

Now we inspect an example of a quadratic form of these parameters under the optimal distribution that appears in the proposed filter and smoother equations. We decompose it using the definition of the quadratic. Note that in the optimal distribution, each row of $\mathbf{A}$ and $\mathbf{B}$ are independent from the other rows.

$$\langle \mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{B} \rangle = \langle \sum_{i=1}^{H} \sum_{j=1}^{H} \mathbf{Q}^{-1}_{(i,j)} \mathbf{a}_{(i)} \mathbf{b}^\mathsf{T}_{(j)} \rangle \qquad \text{Definition of quadratic.} \tag{248}$$

$$= \sum_{i=1}^{H} \sum_{j=1}^{H} \langle \mathbf{Q}^{-1}_{(i,j)} \mathbf{a}_{(i)} \mathbf{b}^\mathsf{T}_{(j)} \rangle \qquad \text{Linearity of expected value.} \tag{249}$$

Since the noise covariance is diagonal, the summand is non-zero only when $i = j = h$:

$$\langle \mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{B} \rangle = \sum_{h=1}^{H} \langle \tau_h \mathbf{a}_{(h)} \mathbf{b}^\mathsf{T}_{(h)} \rangle. \tag{250}$$

Using standard results of the Normal-Gamma distribution [7], this expected value is

$$\langle \mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{B} \rangle = \sum_{h=1}^{H} \langle \tau_h \rangle (\langle \mathbf{a}_{(h)} \rangle \langle \mathbf{b}_{(h)} \rangle^\mathsf{T} + \text{cov}[\mathbf{a}_{(h)}, \mathbf{b}^\mathsf{T}_{(h)}]) \tag{251}$$

$$= \sum_{h=1}^{H} \langle \tau_h \rangle (\langle \mathbf{a}_{(h)} \rangle \langle \mathbf{b}_{(h)} \rangle^\mathsf{T} + \sum_{h=1}^{H} \langle \tau_h \rangle \text{cov}[\mathbf{a}_{(h)}, \mathbf{b}^\mathsf{T}_{(h)}] \tag{252}$$

$$= \langle \mathbf{A} \rangle^\mathsf{T} \text{diag}(\langle \boldsymbol{\tau} \rangle) \langle \mathbf{B} \rangle + \sum_{h=1}^{H} \langle \tau_h \rangle \text{cov}[\mathbf{a}_{(h)}, \mathbf{b}^\mathsf{T}_{(h)}]. \tag{253}$$

Finally, since $\langle \mathbf{Q}^{-1} \rangle = \text{diag}(\langle \boldsymbol{\tau} \rangle)$, we substitute in the covariance from the optimal distribution and simplify:

$$\langle \mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{B} \rangle = \langle \mathbf{A} \rangle^\mathsf{T} \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{B} \rangle + \sum_{h=1}^{H} \langle \tau_h \rangle \text{cov}[\mathbf{a}_{(h)}, \mathbf{b}^\mathsf{T}_{(h)}] \tag{254}$$

$$= \langle \mathbf{A} \rangle^\mathsf{T} \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{B} \rangle + \sum_{h=1}^{H} \langle \tau_h \rangle \boldsymbol{\Sigma}^h_{ab} \langle \tau_h \rangle^{-1} \tag{255}$$

$$= \langle \mathbf{A} \rangle^\mathsf{T} \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{B} \rangle + \sum_{h=1}^{H} \boldsymbol{\Sigma}^h_{ab} \tag{256}$$

$$= \langle \mathbf{A} \rangle^\mathsf{T} \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{B} \rangle + \boldsymbol{\Sigma}_{AQB}. \tag{257}$$

where $\boldsymbol{\Sigma}_{AQB} = \sum_{h=1}^{H} \boldsymbol{\Sigma}^h_{ab}$. An identical result is obtained by starting from the general expression in equation (233).

D.3. Case 2: Non-conjugate prior structure

More generally, we can assume any prior $p(\boldsymbol{\theta})$ that is not necessarily constructed from a conjugate pair of Normal and Gamma distributions. In this case, the optimal distribution $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ will need to be factorized to keep the variational procedure tractable. Here we assume that only the noise needs to be factored from the dynamics matrices (depending on the prior, it might also be the case where the dynamics matrices need to be factored from one another):

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = q_{\boldsymbol{\theta}}(\mathbf{A}, \mathbf{B}) q_{\boldsymbol{\theta}}(\mathbf{Q}) q_{\boldsymbol{\theta}}(\mathbf{C}, \mathbf{D}) q_{\boldsymbol{\theta}}(\mathbf{R}) \tag{258}$$

Without detailing any particular prior distributions or resulting optimal distribution forms, it is simple to show that the quadratic form of these parameters decomposes as follows. Again, we start with the definition of the quadratic.

$$\begin{aligned}
\langle \mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{B} \rangle &= \langle \sum_{i=1}^{H} \sum_{j=1}^{H} \mathbf{Q}^{-1}_{(i,j)} \mathbf{a}_{(i)} \mathbf{b}_{(j)}^\mathsf{T} \rangle && \text{Definition of quadratic.} && (259) \\
&= \sum_{i=1}^{H} \sum_{j=1}^{H} \langle \mathbf{Q}^{-1}_{(i,j)} \mathbf{a}_{(i)} \mathbf{b}_{(j)}^\mathsf{T} \rangle && \text{Linearity of expected value.} && (260)
\end{aligned}$$

Now, since the optimal distribution of the noise covariance matrix is independent of the other parameters, we get

$$\begin{aligned}
\langle \mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{B} \rangle &= \sum_{i=1}^{H} \sum_{j=1}^{H} \langle \mathbf{Q}^{-1} \rangle_{(i,j)} \langle \mathbf{a}_{(i)} \mathbf{b}_{(j)}^\mathsf{T} \rangle && \mathbf{Q} \text{ is independent of } \mathbf{A} \text{ and } \mathbf{B}. && (261) \\
&= \sum_{i=1}^{H} \sum_{j=1}^{H} \langle \mathbf{Q}^{-1} \rangle_{(i,j)} \left(\langle \mathbf{a}_{(i)} \rangle \langle \mathbf{b}_{(j)} \rangle^\mathsf{T} + \mathrm{cov}[\mathbf{a}_{(i)}, \mathbf{b}_{(j)}] \right) && \text{Expand the second moment.} && (262) \\
&= \langle \mathbf{A} \rangle^\mathsf{T} \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{B} \rangle + \sum_{i=1}^{H} \sum_{j=1}^{H} \langle \mathbf{Q}^{-1} \rangle_{(i,j)} \mathrm{cov}[\mathbf{a}_{(i)}, \mathbf{b}_{(j)}] && \text{Simplify first term with definition of quadratic.} && (263)
\end{aligned}$$

In summary, the quadratic forms of the random parameters under their optimal distributions can be expressed as

$$\begin{aligned}
\langle \mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{A} \rangle &= \langle \mathbf{A} \rangle^\mathsf{T} \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{A} \rangle + \boldsymbol{\Sigma}_{AQA} && (264) \\
\langle \mathbf{A}^\mathsf{T} \mathbf{Q}^{-1} \mathbf{B} \rangle &= \langle \mathbf{A} \rangle^\mathsf{T} \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{B} \rangle + \boldsymbol{\Sigma}_{AQB} && (265) \\
\langle \mathbf{C}^\mathsf{T} \mathbf{R}^{-1} \mathbf{C} \rangle &= \langle \mathbf{C} \rangle^\mathsf{T} \langle \mathbf{R}^{-1} \rangle \langle \mathbf{C} \rangle + \boldsymbol{\Sigma}_{CRC} && (266) \\
\langle \mathbf{C}^\mathsf{T} \mathbf{R}^{-1} \mathbf{D} \rangle &= \langle \mathbf{C} \rangle^\mathsf{T} \langle \mathbf{R}^{-1} \rangle \langle \mathbf{D} \rangle + \boldsymbol{\Sigma}_{CRD} && (267)
\end{aligned}$$

where we have defined

$$\begin{aligned}
\boldsymbol{\Sigma}_{AQA} &= \sum_{i=1}^{H} \sum_{j=1}^{H} \langle \mathbf{Q}^{-1} \rangle_{(i,j)} \mathrm{cov}[\mathbf{a}_{(i)}, \mathbf{a}_{(j)}] && (268) \\
\boldsymbol{\Sigma}_{AQB} &= \sum_{i=1}^{H} \sum_{j=1}^{H} \langle \mathbf{Q}^{-1} \rangle_{(i,j)} \mathrm{cov}[\mathbf{a}_{(i)}, \mathbf{b}_{(j)}] && (269) \\
\boldsymbol{\Sigma}_{CRC} &= \sum_{i=1}^{H} \sum_{j=1}^{V} \langle \mathbf{R}^{-1} \rangle_{(i,j)} \mathrm{cov}[\mathbf{c}_{(i)}, \mathbf{c}_{(j)}] && (270) \\
\boldsymbol{\Sigma}_{CRD} &= \sum_{i=1}^{H} \sum_{j=1}^{V} \langle \mathbf{R}^{-1} \rangle_{(i,j)} \mathrm{cov}[\mathbf{c}_{(i)}, \mathbf{d}_{(j)}]. && (271)
\end{aligned}$$

These have the same form as the general expression given in equation (233).

E. SCHUR COMPLEMENTS

The Schur complements were used extensively to derive the forward-backward equations. They are given here for reference.

$$\begin{pmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} \tag{272}$$

$$\mathbf{\Sigma}_{11} = \left(\mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}\right)^{-1} \tag{273}$$

$$= \mathbf{\Lambda}_{11}^{-1} + \mathbf{\Lambda}_{11}^{-1}\mathbf{\Lambda}_{12}\mathbf{\Sigma}_{22}\mathbf{\Lambda}_{21}\mathbf{\Lambda}_{11}^{-1} \tag{274}$$

$$\mathbf{\Sigma}_{22} = \left(\mathbf{\Lambda}_{22} - \mathbf{\Lambda}_{21}\mathbf{\Lambda}_{11}^{-1}\mathbf{\Lambda}_{12}\right)^{-1} \tag{275}$$

$$= \mathbf{\Lambda}_{22}^{-1} + \mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}\mathbf{\Sigma}_{11}\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1} \tag{276}$$

$$\mathbf{\Sigma}_{12} = -\mathbf{\Lambda}_{11}^{-1}\mathbf{\Lambda}_{12}\mathbf{\Sigma}_{22} \tag{277}$$

$$= -\mathbf{\Sigma}_{11}\mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1} \tag{278}$$

$$\mathbf{\Sigma}_{21} = -\mathbf{\Sigma}_{22}\mathbf{\Lambda}_{21}\mathbf{\Lambda}_{11}^{-1} \tag{279}$$

$$= -\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}\mathbf{\Sigma}_{11} \tag{280}$$

F. REFERENCES

[1] C. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, 2006.

[2] H. Rauch, F. Tung, and C. Striebel, "Maximum likelihood estimates of linear dynamical systems," *AIAA Journal*, vol. 3, no. 8, pp. 1445–1450, Aug. 1965.

[3] A. Mathai and S. Provost, *Quadratic forms in random variables: Theory and applications*, Marcel Dekker, 1992.

[4] M. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, University College London, May 2003.

[5] D. Barber, A. Cemgil, and S. Chiappa, Eds., *Bayesian Time Series Models*, Cambridge University Press, 2011.

[6] J. Luttinen, "Fast variational Bayesian linear state-space model," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, Berlin, Heidelberg, 2013, pp. 305–320.

[7] J. Bernardo and A. Smith, *Bayesian theory*, John Wiley & Sons, Inc., 2000.