Probabilistic Derivation of the Basic Kalman Filter & Smoother*

Aaron A. D'Souza
adsouza@usc.edu

1 Introduction

Consider a system having a state $\boldsymbol{\theta}_k$ at time step k which is observed by some process generating an observation $\mathbf{z}_k$. The state transition and observation equations can be written as follows:

$$\begin{aligned}\boldsymbol{\theta}_k &= \mathbf{f}\left(\boldsymbol{\theta}_{k-1}, \mathbf{w}_k\right)\\ \mathbf{z}_k &= \mathbf{g}\left(\boldsymbol{\theta}_k, \mathbf{v}_k\right)\end{aligned}$$

where $\mathbf{w}_k$ and $\mathbf{v}_k$ are system and observation noise respectively. Given a set of observations $\mathcal{D}_k = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\}$, we wish to determine $p(\boldsymbol{\theta}_k|\mathcal{D}_k)$, the distribution over the state at the current time. Using Bayes rule we can write the following:

$$\begin{aligned}p(\boldsymbol{\theta}_k|\mathcal{D}_k) &= p(\boldsymbol{\theta}_k|\mathbf{z}_k, \mathcal{D}_{k-1}) &(1)\\ &\propto p(\mathbf{z}_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1}) &(2)\end{aligned}$$

Also, given the Markov structure of the problem, we have:

$$p(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k-1}|\mathcal{D}_{k-1}) = p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})p(\boldsymbol{\theta}_{k-1}|\mathcal{D}_{k-1}) \tag{3}$$

From Eqs. (2) and (3) we can derive the following recursive state estimaton equations:

$$\left.\begin{aligned}p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1}) = \int p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})p(\boldsymbol{\theta}_{k-1}|\mathcal{D}_{k-1})d\boldsymbol{\theta}_{k-1}\end{aligned}\right\} \quad \text{Prediction} \tag{4}$$

$$\left.\begin{aligned}p(\boldsymbol{\theta}_k|\mathcal{D}_k) &= c_k p(\mathbf{z}_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1})\\ c_k &= \int p(\mathbf{z}_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1})\end{aligned}\right\} \quad \text{Filter update} \tag{5}$$

In general these equations are difficult to evaluate. For special cases of state transition and observation functions, and noise distributions however, these equations become exactly solvable.

*This derivation is pretty detailed, and might have a typo or two. If you find any, please let me know!

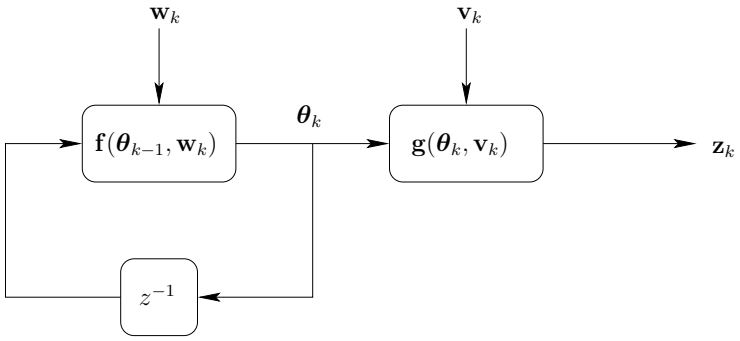

Figure 1: Generic system diagram.

2 Linear-Gaussian state space model

The Kalman filter [2] essentially solves the prediction and filter update equations for the special case in which the system transition and observation functions are linear in the state and noise, and the noise is Gaussian. In this case since the noise is Gaussian, the integrals are analytically tractable, and the linearity of the transition and observation functions ensure that the state and observation distributions retain their Gaussian form at each step. We can write the state transition and observation equations as follows:

$$\boldsymbol{\theta}_k = \mathbf{F}\boldsymbol{\theta}_{k-1} + \boldsymbol{\Phi}\mathbf{w}_k \tag{6}$$
$$\mathbf{z}_k = \mathbf{G}\boldsymbol{\theta}_k + \boldsymbol{\Psi}\mathbf{v}_k \tag{7}$$

where we assume $\mathbf{w}_k \sim \text{Normal}\,(\mathbf{w}_k; \mathbf{0}, \mathbf{Q})$, $\mathbf{v}_k \sim \text{Normal}\,(\mathbf{v}_k; \mathbf{0}, \mathbf{R})$. We define the following statistics for the prior and posterior distributions:

$$\left.\begin{aligned} \mathbf{a}_{k|k-1} &= \langle \boldsymbol{\theta}_k | \mathcal{D}_{k-1} \rangle \\ \mathbf{P}_{k|k-1} &= \text{Cov}(\boldsymbol{\theta}_k | \mathcal{D}_{k-1}) \end{aligned}\right\} \text{ Prior mean and covariance}$$

$$\left.\begin{aligned} \mathbf{a}_{k} &= \langle \boldsymbol{\theta}_k | \mathcal{D}_{k} \rangle \\ \mathbf{P}_{k} &= \text{Cov}(\boldsymbol{\theta}_k | \mathcal{D}_{k}) \end{aligned}\right\} \text{ Posterior mean and covariance}$$

From Eqs. (6) and (7) we can also infer the following:

$$\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1} \sim \text{Normal}\left(\boldsymbol{\theta}_k; \mathbf{F}\boldsymbol{\theta}_{k-1}, \boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right) \tag{8}$$
$$\mathbf{z}_k | \boldsymbol{\theta}_k \sim \text{Normal}\left(\mathbf{z}_k; \mathbf{G}\boldsymbol{\theta}_k, \boldsymbol{\Psi}\mathbf{R}\boldsymbol{\Psi}^T\right) \tag{9}$$

2.1 Prediction

Eq. (4) allows us to create a prior distribution of the current state $\boldsymbol{\theta}_k$ with only knowledge of past observations $\mathcal{D}_{k-1}$. The reason this is a *prior* distribution is because we have yet to make an observation $\mathbf{z}_k$ in this state. Once we make the observation, we will need to integrate it with this distribution to generate a *posterior* distribution.

$$p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1}) = \int p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1})p(\boldsymbol{\theta}_{k-1}|\mathcal{D}_{k-1})d\boldsymbol{\theta}_{k-1} \tag{10}$$

$$= \int \text{Normal}\left(\boldsymbol{\theta}_k; \mathbf{F}\boldsymbol{\theta}_{k-1}, \boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right) \text{Normal}\left(\boldsymbol{\theta}_{k-1}; \mathbf{a}_{k-1}, \mathbf{P}_{k-1}\right) d\boldsymbol{\theta}_{k-1} \tag{11}$$

$$= k \int \exp\left\{-\frac{1}{2}A\right\} d\boldsymbol{\theta}_{k-1} \tag{12}$$

The term A inside the exponent of Eq. (12) can be expanded as follows:

$$\begin{aligned} A &= \left(\boldsymbol{\theta}_k - \mathbf{F}\boldsymbol{\theta}_{k-1}\right)^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \left(\boldsymbol{\theta}_k - \mathbf{F}\boldsymbol{\theta}_{k-1}\right) + \left(\boldsymbol{\theta}_{k-1} - \mathbf{a}_{k-1}\right)^T \mathbf{P}_{k-1}^{-1} \left(\boldsymbol{\theta}_{k-1} - \mathbf{a}_{k-1}\right) \\ &= {\boldsymbol{\theta}_{k-1}}^T \underbrace{\left(\mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F} + \mathbf{P}_{k-1}^{-1}\right)}_{\boldsymbol{\Delta}_{k-1}} \boldsymbol{\theta}_{k-1} - 2\boldsymbol{\theta}_{k-1}^T \underbrace{\left(\mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \boldsymbol{\theta}_k + \mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1}\right)}_{\mathbf{C}} \\ &\quad + {\boldsymbol{\theta}_k}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \boldsymbol{\theta}_k + {\mathbf{a}_{k-1}}^T \mathbf{P}_{k-1}^{-1} \mathbf{a}_{k-1} \end{aligned}$$

This is a quadratic in $\boldsymbol{\theta}_{k-1}$, which can be written as

$$A = \left(\boldsymbol{\theta}_{k-1} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\theta}_{k-1} - \boldsymbol{\mu}\right) - \mathbf{C}^T \boldsymbol{\Delta}_{k-1}^{-1} \mathbf{C} + {\boldsymbol{\theta}_k}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \boldsymbol{\theta}_k + {\mathbf{a}_{k-1}}^T \mathbf{P}_{k-1}^{-1} \mathbf{a}_{k-1}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\Delta}_{k-1}^{-1}$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{C}$. Substituting this result back into the exponent of Eq. (12) we get:

$$p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1}) = k \int \exp\left\{-\frac{1}{2}A\right\} d\boldsymbol{\theta}_{k-1} \tag{13}$$

$$\begin{aligned} &= k \exp\left\{-\frac{1}{2}\left[{\boldsymbol{\theta}_k}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \boldsymbol{\theta}_k + {\mathbf{a}_{k-1}}^T \mathbf{P}_{k-1}^{-1} \mathbf{a}_{k-1} - \mathbf{C}^T \boldsymbol{\Delta}_{k-1}^{-1} \mathbf{C}\right]\right\} \\ &\quad \cdot \int \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}_{k-1} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\theta}_{k-1} - \boldsymbol{\mu}\right)\right\} d\boldsymbol{\theta}_{k-1} \end{aligned} \tag{14}$$

$$= k \left(2\pi\right)^{d/2} \left|\boldsymbol{\Sigma}\right|^{1/2} \exp\left\{-\frac{1}{2} \underbrace{\left[{\boldsymbol{\theta}_k}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \boldsymbol{\theta}_k + {\mathbf{a}_{k-1}}^T \mathbf{P}_{k-1}^{-1} \mathbf{a}_{k-1} - \mathbf{C}^T \boldsymbol{\Delta}_{k-1}^{-1} \mathbf{C}\right]}_{D}\right\} \tag{15}$$

This distribution has a quadratic inside an exponential term implying that $p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1})$ has a Gaussian form. We can expand the quadratic term D as follows:

$$D = -\left(\mathbf{F}^T\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\boldsymbol{\theta}_k + \mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1}\right)^T \mathbf{\Delta}_{k-1}^{-1}\left(\mathbf{F}^T\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\boldsymbol{\theta}_k + \mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1}\right) + {\boldsymbol{\theta}_k}^T\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\boldsymbol{\theta}_k + {\mathbf{a}_{k-1}}^T\mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1} \tag{16}$$

Consider terms in D that are quadratic in $\boldsymbol{\theta}_k$:

$$\begin{aligned} D_2 &= {\boldsymbol{\theta}_k}^T\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\boldsymbol{\theta}_k - {\boldsymbol{\theta}_k}^T\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\mathbf{F}\mathbf{\Delta}_{k-1}^{-1}\mathbf{F}^T\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\boldsymbol{\theta}_k \\ &= {\boldsymbol{\theta}_k}^T\left[\left(\mathbf{\Phi Q \Phi}^T\right)^{-1} - \left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\mathbf{F}\left(\mathbf{F}^T\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\mathbf{F} + \mathbf{P}_{k-1}^{-1}\right)^{-1}\mathbf{F}^T\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\right]\boldsymbol{\theta}_k \qquad (17) \\ &= {\boldsymbol{\theta}_k}^T\left(\mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T + \mathbf{\Phi Q \Phi}^T\right)^{-1}\boldsymbol{\theta}_k \qquad (18) \end{aligned}$$

Since this is the only term in D that is quadratic in $\boldsymbol{\theta}_k$, we can infer the covariance of the distribution as follows:

$$\Rightarrow \boxed{\mathbf{P}_{k|k-1} = \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T + \mathbf{\Phi Q \Phi}^T} \tag{19}$$

Note that Eq. (18), results from using the matrix inversion lemma (Appendix A.1). To derive the mean, we consider terms in D that are linear in $\boldsymbol{\theta}_k$

$$\begin{aligned} D_1 &= -2\boldsymbol{\theta}_k\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\mathbf{F}\left(\mathbf{F}^T\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\mathbf{F} + \mathbf{P}_{k-1}^{-1}\right)^{-1}\mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1} \qquad (20) \\ &= -2\boldsymbol{\theta}_k\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\mathbf{F}\left[\mathbf{P}_{k-1} - \mathbf{P}_{k-1}\mathbf{F}^T\left(\mathbf{\Phi Q \Phi}^T + \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T\right)^{-1}\mathbf{F}\mathbf{P}_{k-1}\right]\mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1} \qquad (21) \\ &= -2\boldsymbol{\theta}_k\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\left[\mathbf{F} - \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T\left(\mathbf{\Phi Q \Phi}^T + \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T\right)^{-1}\mathbf{F}\right]\mathbf{a}_{k-1} \\ &= -2\boldsymbol{\theta}_k\left(\mathbf{\Phi Q \Phi}^T\right)^{-1}\begin{bmatrix} \mathbf{F} - \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T\left(\mathbf{\Phi Q \Phi}^T + \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T\right)^{-1}\mathbf{F} \\ -\mathbf{\Phi Q \Phi}^T\left(\mathbf{\Phi Q \Phi}^T + \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T\right)^{-1}\mathbf{F} \\ +\mathbf{\Phi Q \Phi}^T\left(\mathbf{\Phi Q \Phi}^T + \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T\right)^{-1}\mathbf{F} \end{bmatrix}\mathbf{a}_{k-1} \\ &= -2\boldsymbol{\theta}_k\left(\mathbf{\Phi Q \Phi}^T + \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T\right)^{-1}\mathbf{F}\mathbf{a}_{k-1} \\ &= -2\boldsymbol{\theta}_k\mathbf{P}_{k|k-1}^{-1}\mathbf{F}\mathbf{a}_{k-1} \\ &\Rightarrow \boxed{\mathbf{a}_{k|k-1} = \mathbf{F}\mathbf{a}_{k-1}} \qquad (22) \end{aligned}$$

Where again, Eq. (21) uses the matrix inversion lemma (Appendix A.1). Hence, the prior distribution $p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1})$ is Gaussian:

$$p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1}) = \text{Normal}\left(\boldsymbol{\theta}_k; \mathbf{a}_{k|k-1}, \mathbf{P}_{k|k-1}\right) \tag{23}$$

$$\mathbf{a}_{k|k-1} = \mathbf{F}\mathbf{a}_{k-1} \tag{24}$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}\mathbf{P}_{k-1}\mathbf{F}^T + \boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T \tag{25}$$

2.2 Filter Update

Now that we have the prior distribution $p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1})$ at the current timestep, we wish to integrate the current observation $\mathbf{z}_k$ with this prior to generate a posterior distribution $p(\boldsymbol{\theta}_k|\mathcal{D}_k)$. From Eq. (2)

$$p(\boldsymbol{\theta}_k|\mathcal{D}_k) \propto p(\mathbf{z}_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\mathcal{D}_{k-1})$$

which, substituting from Eqs. (9) and (23), we get:

$$\propto \exp\left\{ \underbrace{-\frac{1}{2}\left[\begin{array}{l} \left(\mathbf{z}_k - \mathbf{G}\boldsymbol{\theta}_k\right)^T \left(\boldsymbol{\Psi}\mathbf{R}\boldsymbol{\Psi}^T\right)^{-1} \left(\mathbf{z}_k - \mathbf{G}\boldsymbol{\theta}_k\right) \\ +\left(\boldsymbol{\theta}_k - \mathbf{a}_{k|k-1}\right)^T \mathbf{P}_{k|k-1}^{-1} \left(\boldsymbol{\theta}_k - \mathbf{a}_{k|k-1}\right) \end{array} \right]}_{E} \right\}$$

Now:

$$\begin{aligned} E &= \left(\mathbf{z}_k - \mathbf{G}\boldsymbol{\theta}_k\right)^T \left(\boldsymbol{\Psi}\mathbf{R}\boldsymbol{\Psi}^T\right)^{-1} \left(\mathbf{z}_k - \mathbf{G}\boldsymbol{\theta}_k\right) + \left(\boldsymbol{\theta}_k - \mathbf{a}_{k|k-1}\right)^T \mathbf{P}_{k|k-1}^{-1} \left(\boldsymbol{\theta}_k - \mathbf{a}_{k|k-1}\right) \\ &= {\boldsymbol{\theta}_k}^T \left(\mathbf{G}^T \left(\boldsymbol{\Psi}\mathbf{R}\boldsymbol{\Psi}^T\right)^{-1} \mathbf{G} + \mathbf{P}_{k|k-1}^{-1}\right) \boldsymbol{\theta}_k \\ &\quad - 2\boldsymbol{\theta}_k^T \left(\mathbf{G}^T \left(\boldsymbol{\Psi}\mathbf{R}\boldsymbol{\Psi}^T\right)^{-1} \mathbf{z}_k + \mathbf{P}_{k|k-1}^{-1}\mathbf{a}_{k|k-1}\right) + const_{\boldsymbol{\theta}_k} \end{aligned} \tag{26}$$

From the quadratic part of Eq. (26), we can deduce:

$$\mathbf{P}_k = \left(\mathbf{G}^T \left(\boldsymbol{\Psi}\mathbf{R}\boldsymbol{\Psi}^T\right)^{-1} \mathbf{G} + \mathbf{P}_{k|k-1}^{-1}\right)^{-1} \tag{27}$$

$$= \mathbf{P}_{k|k-1} - \mathbf{P}_{k|k-1}\mathbf{G}^T \left(\boldsymbol{\Psi}\mathbf{R}\boldsymbol{\Psi}^T + \mathbf{G}^T\mathbf{P}_{k|k-1}\mathbf{G}\right)^{-1} \mathbf{G}\mathbf{P}_{k|k-1} \tag{28}$$

$$= \mathbf{P}_{k|k-1} - \mathbf{P}_{k|k-1}\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G}\mathbf{P}_{k|k-1} \qquad \text{where } \mathbf{K} = \boldsymbol{\Psi}\mathbf{R}\boldsymbol{\Psi}^T + \mathbf{G}^T\mathbf{P}_{k|k-1}\mathbf{G} \tag{29}$$

Where again, Eq. (28) uses the matrix inversion lemma (Appendix A.1). Also, from the linear

part of Eq. (26), we can deduce:

$$\begin{aligned}\mathbf{a}_k &= \mathbf{P}_k \left(\mathbf{G}^T \left(\mathbf{\Psi R \Psi}^T \right)^{-1} \mathbf{z}_k + \mathbf{P}_{k|k-1}^{-1} \mathbf{a}_{k|k-1} \right) \\ &= \left(\mathbf{P}_{k|k-1} - \mathbf{P}_{k|k-1} \mathbf{G}^T \mathbf{K}^{-1} \mathbf{G} \mathbf{P}_{k|k-1} \right) \left(\mathbf{G}^T \left(\mathbf{\Psi R \Psi}^T \right)^{-1} \mathbf{z}_k + \mathbf{P}_{k|k-1}^{-1} \mathbf{a}_{k|k-1} \right) \\ &= \mathbf{a}_{k|k-1} - \mathbf{P}_{k|k-1} \mathbf{G}^T \mathbf{K}^{-1} \mathbf{G} \mathbf{a}_{k|k-1} + \mathbf{P}_{k|k-1} \mathbf{G}^T \underbrace{\left[\mathbf{I} - \mathbf{K}^{-1} \mathbf{G} \mathbf{P}_{k|k-1} \mathbf{G}^T \right]}_{F} \left(\mathbf{\Psi R \Psi}^T \right)^{-1} \mathbf{z}_k \end{aligned} \tag{30}$$

Now:

$$\begin{aligned} F &= \mathbf{I} - \mathbf{K}^{-1} \mathbf{G} \mathbf{P}_{k|k-1} \mathbf{G}^T \\ &= \mathbf{I} - \mathbf{K}^{-1} \mathbf{G} \mathbf{P}_{k|k-1} \mathbf{G}^T - \mathbf{K}^{-1} \mathbf{\Psi R \Psi}^T + \mathbf{K}^{-1} \mathbf{\Psi R \Psi}^T \\ &= \mathbf{K}^{-1} \mathbf{\Psi R \Psi}^T \end{aligned}$$

Hence

$$\mathbf{a}_k = \mathbf{a}_{k|k-1} + \mathbf{P}_{k|k-1} \mathbf{G}^T \mathbf{K}^{-1} \left(\mathbf{z}_k - \mathbf{G} \mathbf{a}_{k|k-1} \right)$$

This allows us to characterize the posterior distribution $p(\boldsymbol{\theta}_k | \mathcal{D}_k)$ as a Gaussian with the following mean and variance:

$$\begin{aligned} \mathbf{a}_k &= \mathbf{a}_{k|k-1} + \mathbf{P}_{k|k-1} \mathbf{G}^T \mathbf{K}^{-1} \left(\mathbf{z}_k - \mathbf{G} \mathbf{a}_{k|k-1} \right) \\ \mathbf{P}_k &= \mathbf{P}_{k|k-1} - \mathbf{P}_{k|k-1} \mathbf{G}^T \mathbf{K}^{-1} \mathbf{G} \mathbf{P}_{k|k-1} \end{aligned}$$

where

$$\mathbf{K} = \mathbf{\Psi R \Psi}^T + \mathbf{G}^T \mathbf{P}_{k|k-1} \mathbf{G}$$

Certain applications (in particular, Bayesian inference over the state-transition, and observation matrices $\mathbf{F}$ and $\mathbf{G}$), preclude the use of the matrix inversion lemma, since expectations are not propagated correctly through this operation [1]. If we choose to do without this useful simplification, then we can rewrite our equations as follows:

For the prediction equations, we use Eqs. (17) and (20):

$$\mathbf{P}_{k|k-1} = \left[\left(\mathbf{\Phi Q \Phi}^T \right)^{-1} - \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \mathbf{F} \left(\mathbf{F}^T \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \mathbf{F} + \mathbf{P}_{k-1}^{-1} \right)^{-1} \mathbf{F}^T \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \right]^{-1} \tag{31}$$

$$\mathbf{a}_{k|k-1} = \mathbf{P}_{k|k-1} \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \mathbf{F} \left(\mathbf{F}^T \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \mathbf{F} + \mathbf{P}_{k-1}^{-1} \right)^{-1} \mathbf{P}_{k-1}^{-1} \mathbf{a}_{k-1} \tag{32}$$

And for the filter update equations, we use Eqs. (27) and (30):

$$\mathbf{P}_k = \left(\mathbf{G}^T \left(\mathbf{\Psi R \Psi}^T \right)^{-1} \mathbf{G} + \mathbf{P}_{k|k-1}^{-1} \right)^{-1} \tag{33}$$

$$\mathbf{a}_k = \mathbf{P}_k \left(\mathbf{G}^T \left(\mathbf{\Psi R \Psi}^T \right)^{-1} \mathbf{z}_k + \mathbf{P}_{k|k-1}^{-1} \mathbf{a}_{k|k-1} \right) \tag{34}$$

Substituting the prediction equations into the filter update equations, we can get a one-stage, recursion at each timestep (provided we have already observed $\mathbf{z}_k$):

$$\mathbf{P}_k = \begin{bmatrix} \mathbf{G}^T \left(\mathbf{\Psi R \Psi}^T \right)^{-1} \mathbf{G} + \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \\ - \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \mathbf{F} \left(\mathbf{F}^T \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \mathbf{F} + \mathbf{P}_{k-1}^{-1} \right)^{-1} \mathbf{F}^T \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \end{bmatrix}^{-1} \tag{35}$$

$$\mathbf{a}_k = \mathbf{P}_k \left[\mathbf{G}^T \left(\mathbf{\Psi R \Psi}^T \right)^{-1} \mathbf{z}_k + \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \mathbf{F} \left(\mathbf{F}^T \left(\mathbf{\Phi Q \Phi}^T \right)^{-1} \mathbf{F} + \mathbf{P}_{k-1}^{-1} \right)^{-1} \mathbf{P}_{k-1}^{-1} \mathbf{a}_{k-1} \right] \tag{36}$$

3 The Kalman Smoother

(NOTE: This derivation follows the pattern of [1])

Given a set of observations $\mathcal{D}_K$, the Kalman *filter* detailed in the preceding sections effectively computes the quantity $p(\boldsymbol{\theta}_K|\mathcal{D}_K)$, i.e., the distribution of the *last* state in the sequence, taking into account all observations up to that timestep. This is accomplished by performing the following recursion *forward* in time (obtained by computing the prediction and filter update equations):

$$p(\boldsymbol{\theta}_k|\mathcal{D}_k) \propto p(\mathbf{z}_k|\boldsymbol{\theta}_k) \int p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}) p(\boldsymbol{\theta}_{k-1}|\mathcal{D}_{k-1}) d\boldsymbol{\theta}_{k-1} \tag{37}$$

Starting with an initial Gaussian distribution $p(\boldsymbol{\theta}_0) = \text{Normal}\,(\boldsymbol{\theta}_0; \mathbf{a}_0, \mathbf{P}_0)$, the filtered distributions at all subsequent timesteps are also Gaussian with means and variances given by Eqs. (35) and (36)

It is important to note that for any timestep k, the filtered state estimate $p(\boldsymbol{\theta}_k|\mathcal{D}_k)$ takes into account all information *up to that time* k, and not after. Hence the process of filtering ignores the effect of data observed *after* k, i.e. at timesteps $k+1, \ldots, K$, in estimating the distribution of $\boldsymbol{\theta}_k$.

If we wish to propagate this (in some sense) *future* information in the anti-causal direction, and determine the influence on *any* state $\boldsymbol{\theta}_k$ given all observations past *and* future to that timestep k, i.e., compute $p(\boldsymbol{\theta}_k|\mathcal{D}_K)$, where $k < K$, then we need to derive the following recursion:

$$
\begin{aligned}
p(\boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_k | \mathcal{D}_K) &= p(\boldsymbol{\theta}_{k-1} | \boldsymbol{\theta}_k, \mathcal{D}_K) p(\boldsymbol{\theta}_k | \mathcal{D}_K) \\
&= p(\boldsymbol{\theta}_{k-1} | \boldsymbol{\theta}_k, \mathcal{D}_{k-1}) p(\boldsymbol{\theta}_k | \mathcal{D}_K) \\
&= \frac{p(\boldsymbol{\theta}_{k-1} | \boldsymbol{\theta}_k, \mathcal{D}_{k-1}) p(\boldsymbol{\theta}_k | \mathcal{D}_{k-1})}{p(\boldsymbol{\theta}_k | \mathcal{D}_{k-1})} p(\boldsymbol{\theta}_k | \mathcal{D}_K) \\
&= \frac{p(\boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_k | \mathcal{D}_{k-1})}{\int p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}) p(\boldsymbol{\theta}_{k-1} | \mathcal{D}_{k-1}) d\boldsymbol{\theta}_{k-1}} p(\boldsymbol{\theta}_k | \mathcal{D}_K) \\
&= \frac{p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}) p(\boldsymbol{\theta}_{k-1} | \mathcal{D}_{k-1})}{\int p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}) p(\boldsymbol{\theta}_{k-1} | \mathcal{D}_{k-1}) d\boldsymbol{\theta}_{k-1}} p(\boldsymbol{\theta}_k | \mathcal{D}_K)
\end{aligned}
$$

Integrating over $\boldsymbol{\theta}_k$ gives us the required backward (in time) recursion:

$$
p(\boldsymbol{\theta}_{k-1} | \mathcal{D}_K) = \int \left[\frac{p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}) p(\boldsymbol{\theta}_{k-1} | \mathcal{D}_{k-1})}{\int p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}) p(\boldsymbol{\theta}_{k-1} | \mathcal{D}_{k-1}) d\boldsymbol{\theta}_{k-1}} \right] p(\boldsymbol{\theta}_k | \mathcal{D}_K) d\boldsymbol{\theta}_k \tag{38}
$$

Note that in this equation, the distributions $p(\boldsymbol{\theta}_{k-1} | \mathcal{D}_{k-1})$ are already available as our *filtered* estimates for each timestep. It is clear that for the final timestep K, the *filtered* and *smoothed* distributions are identical, i.e. $p(\boldsymbol{\theta}_K | \mathcal{D}_K)$, and hence after the forward (filtering) recursion, we are provided with a starting point for the backward recursion. Assume that for the kth timestep our smoothed estimate is represented by the following distribution:

$$
\boldsymbol{\theta}_k | \mathcal{D}_K \sim \text{Normal}\left(\boldsymbol{\theta}_k; \boldsymbol{\eta}_k, \boldsymbol{\Lambda}_k\right) \tag{39}
$$

The smaller integral in the denominator of the integrand of Eq. (38) is exactly the prediction integral of the Kalman filter in Eq. (10), which from Eq. (15) is shown to be the following:

$$
\begin{aligned}
\int p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}) p(\boldsymbol{\theta}_{k-1} | \mathcal{D}_{k-1}) d\boldsymbol{\theta}_{k-1} &= p(\boldsymbol{\theta}_k | \mathcal{D}_{k-1}) && (40) \\
&\propto \exp\left\{ -\frac{1}{2} D \right\} && (41)
\end{aligned}
$$

where D is expanded as (repeated from Eq. (16) for convenience):

$$
\begin{aligned}
D = -\Big(\mathbf{F}^T \left(\boldsymbol{\Phi} \mathbf{Q} \boldsymbol{\Phi}^T \right)^{-1} \boldsymbol{\theta}_k + \mathbf{P}_{k-1}^{-1} \mathbf{a}_{k-1} \Big)^T \boldsymbol{\Delta}_{k-1}^{-1} \Big(\mathbf{F}^T \left(\boldsymbol{\Phi} \mathbf{Q} \boldsymbol{\Phi}^T \right)^{-1} \boldsymbol{\theta}_k + \mathbf{P}_{k-1}^{-1} \mathbf{a}_{k-1} \Big) \\
+ {\boldsymbol{\theta}_k}^T \left(\boldsymbol{\Phi} \mathbf{Q} \boldsymbol{\Phi}^T \right)^{-1} \boldsymbol{\theta}_k + {\mathbf{a}_{k-1}}^T \mathbf{P}_{k-1}^{-1} \mathbf{a}_{k-1}
\end{aligned} \tag{42}
$$

We can thus write the integral as follows:

$$
\gamma_{k-1}(\boldsymbol{\theta}_{k-1}) \propto \int \exp\left\{ -\frac{1}{2} E \right\} d\boldsymbol{\theta}_k \tag{43}
$$

where:

$$
\begin{aligned}
E = &(\boldsymbol{\theta}_k - \mathbf{F}\boldsymbol{\theta}_{k-1})^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} (\boldsymbol{\theta}_k - \mathbf{F}\boldsymbol{\theta}_{k-1}) + (\boldsymbol{\theta}_{k-1} - \mathbf{a}_{k-1})^T \mathbf{P}_{k-1}^{-1} (\boldsymbol{\theta}_{k-1} - \mathbf{a}_{k-1}) \\
&+ (\boldsymbol{\theta}_k - \boldsymbol{\eta}_k)^T \boldsymbol{\Lambda}_k^{-1} (\boldsymbol{\theta}_k - \boldsymbol{\eta}_k) - {\boldsymbol{\theta}_k}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \boldsymbol{\theta}_k \qquad (44) \\
&+ \left(\mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \boldsymbol{\theta}_k + \mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1}\right)^T \boldsymbol{\Delta}_{k-1}^{-1} \left(\mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \boldsymbol{\theta}_k + \mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1}\right) \\
= &{\boldsymbol{\theta}_k}^T \underbrace{\left(\boldsymbol{\Lambda}_k^{-1} + \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\boldsymbol{\Delta}_{k-1}^{-1}\mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1}\right)}_{\mathbf{B}} \boldsymbol{\theta}_k \\
&- 2\boldsymbol{\theta}_k^T \underbrace{\left(\left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\boldsymbol{\theta}_{k-1} + \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\eta}_k - \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\boldsymbol{\Delta}_{k-1}^{-1}\mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1}\right)}_{\mathbf{C}} \qquad (45) \\
&+ {\boldsymbol{\theta}_{k-1}}^T \mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\boldsymbol{\theta}_{k-1} + (\boldsymbol{\theta}_{k-1} - \mathbf{a}_{k-1})^T \mathbf{P}_{k-1}^{-1} (\boldsymbol{\theta}_{k-1} - \mathbf{a}_{k-1}) \\
&+ {\boldsymbol{\eta}_k}^T \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\eta}_k + {\mathbf{a}_{k-1}}^T \mathbf{P}_{k-1}^{-1}\boldsymbol{\Delta}_{k-1}^{-1}\mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1}
\end{aligned}
$$

Complete squares in $\boldsymbol{\theta}_k$

$$
\begin{aligned}
E = &(\boldsymbol{\theta}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_k - \boldsymbol{\mu}) - \mathbf{C}^T\mathbf{B}^{-1}\mathbf{C} \\
&+ {\boldsymbol{\theta}_{k-1}}^T \mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\boldsymbol{\theta}_{k-1} + (\boldsymbol{\theta}_{k-1} - \mathbf{a}_{k-1})^T \mathbf{P}_{k-1}^{-1} (\boldsymbol{\theta}_{k-1} - \mathbf{a}_{k-1}) \qquad (46) \\
&+ {\boldsymbol{\eta}_k}^T \boldsymbol{\Lambda}_k^{-1}\boldsymbol{\eta}_k + {\mathbf{a}_{k-1}}^T \mathbf{P}_{k-1}^{-1}\boldsymbol{\Delta}_{k-1}^{-1}\mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1}
\end{aligned}
$$

Substituting this expression back in the integral for $\gamma_{k-1}(\boldsymbol{\theta}_{k-1})$ we have:

$$
\begin{aligned}
\gamma_{k-1}(\boldsymbol{\theta}_{k-1}) \propto \exp &\left\{ -\frac{1}{2} \underbrace{\left[{\boldsymbol{\theta}_{k-1}}^T \mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\boldsymbol{\theta}_{k-1} + (\boldsymbol{\theta}_{k-1} - \mathbf{a}_{k-1})^T \mathbf{P}_{k-1}^{-1} (\boldsymbol{\theta}_{k-1} - \mathbf{a}_{k-1}) - \mathbf{C}^T\mathbf{B}^{-1}\mathbf{C} \right]}_{D} \right\} \\
&\cdot \int \exp\left\{ -\frac{1}{2} (\boldsymbol{\theta}_k - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\theta}_k - \boldsymbol{\mu}) \right\} d\boldsymbol{\theta}_k
\end{aligned} \qquad (47)
$$

This expression a Normal distribution, the term D inside the exponent can be written as:

$$
\begin{aligned}
D = &{\theta_{k-1}}^T \left(\mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F} + \mathbf{P}_{k-1}^{-1} - \mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{B}^{-1} \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\right) \theta_{k-1} \\
&- 2\boldsymbol{\theta}_{k-1} \left(\mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1} + \mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{B}^{-1} \left(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\eta}_k - \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\boldsymbol{\Delta}_{k-1}^{-1}\mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1}\right)\right) \\
&+ const_{\boldsymbol{\theta}_{k-1}}
\end{aligned} \qquad (48)
$$

We can thus deduce the parameters of the $p(\boldsymbol{\theta}_{k-1}|\mathcal{D}_K)$ distribution as:

$$\boldsymbol{\Lambda}_{k-1} = \left(\boldsymbol{\Delta}_{k-1} - \mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{B}^{-1} \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\right)^{-1} \tag{49}$$

$$\boldsymbol{\eta}_{k-1} = \boldsymbol{\Lambda}_{k-1} \left(\mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1} + \mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{B}^{-1} \left(\boldsymbol{\Lambda}_k^{-1}\boldsymbol{\eta}_k - \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\boldsymbol{\Delta}_{k-1}^{-1}\mathbf{P}_{k-1}^{-1}\mathbf{a}_{k-1}\right)\right) \tag{50}$$

where:

$$\boldsymbol{\Delta}_{k-1} = \mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F} + \mathbf{P}_{k-1}^{-1} \tag{51}$$

$$\mathbf{B} = \boldsymbol{\Lambda}_k^{-1} + \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \mathbf{F}\boldsymbol{\Delta}_{k-1}^{-1}\mathbf{F}^T \left(\boldsymbol{\Phi}\mathbf{Q}\boldsymbol{\Phi}^T\right)^{-1} \tag{52}$$

A Useful matrix algebra results

A.1 Matrix inversion theorem

The famous Sherman-Morrison-Woodbury theorem:

$$(\mathbf{A} + \mathbf{XRY})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{X}(\mathbf{R}^{-1} + \mathbf{YA}^{-1}\mathbf{X})^{-1}\mathbf{YA}^{-1}$$

A.2 Schur's complements

References

[1] Matthew J. Beal and Zoubin Ghahramani. The variational Kalman smoother. Technical Report GCNU TR 2001-003, Gatsby Computational Neuroscience Unit, University College London, 2001.

[2] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME — Journal of Basic Engineering*, 82:35–45, 1960.