Pattern Recognition and Machine Vision: Kalman Filter

Dr. Simon J.D. Prince
Computer Science
University College London
Gower Street, London, WC1E 6BT

1 Mathematics

We will use the notation $G_{\mathbf{x}}(\mu,\Sigma)$ to represent a Gaussian in the variable **x** with mean μ and covariance Σ.

$$G_{\mathbf{x}}(\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-0.5(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right] \tag{1}$$

1.1 Gaussian Lemma #1

The product of two Gaussians is also a Gaussian. To see this, consider that the exponent of each Gaussian is a quadratic in **x**. When we multiply the two Gaussians, we add the exponents. This produces the sum of two quadratics, which is itself another quadratic. In fact, the following relation holds:

$$G_{\mathbf{x}}(\mathbf{a},\mathbf{A})\,G_{\mathbf{x}}(\mathbf{b},\mathbf{B}) \propto G_{\mathbf{x}}\left((\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}),(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\right) \tag{2}$$

Proof:

$$\begin{aligned}
G_{\mathbf{x}}(\mathbf{a},\mathbf{A})\,G_{\mathbf{x}}(\mathbf{b},\mathbf{B}) &= \frac{1}{(2\pi)^{n}|\mathbf{A}|^{1/2}|\mathbf{B}|^{1/2}} \exp\left[-0.5((\mathbf{x}-\mathbf{a})^T\mathbf{A}^{-1}(\mathbf{x}-\mathbf{a})+(\mathbf{x}-\mathbf{b})^T\mathbf{B}^{-1}(\mathbf{x}-\mathbf{b}))\right]\\
&= k\exp\left[-0.5(\mathbf{x}^T(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{x}+\mathbf{x}^T(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\right.\\
&\qquad\qquad \left.+(\mathbf{a}^T\mathbf{A}^{-1}+\mathbf{b}^T\mathbf{B}^{-1})\mathbf{x})\right]
\end{aligned} \tag{3}$$

where the exponential terms that do not depend on **x** have been subsumed into the constant, k. It is clear from the quadratic term, that this can be re-arranged to form a Gaussian with covariance $(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}$. We can complete the square:

$$\begin{aligned}
G_{\mathbf{x}}(\mathbf{a},\mathbf{A})\,G_{\mathbf{x}}(\mathbf{b},\mathbf{B}) &= k_2\exp\left[-0.5(\mathbf{x}^T(\mathbf{A}^{-1}+\mathbf{B}^{-1})\mathbf{x}+\mathbf{x}^T(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\right.\\
&\qquad +(\mathbf{a}^T\mathbf{A}^{-1}+\mathbf{b}^T\mathbf{B}^{-1})\mathbf{x}\\
&\qquad \left.+(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})^T(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}))\right]\\
&= k_2\exp\Big[-0.5\left(\mathbf{x}-(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\right)^T\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)\\
&\qquad \left(\mathbf{x}-(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b})\right)\Big]\\
&= k_2G_{\mathbf{x}}\left((\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}),(\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}\right)
\end{aligned} \tag{4}$$

as required.

1.2 Gaussian Lemma #2

We will also need a second Gaussian relation: Consider a Gaussian in $\mathbf{x}$ with a mean that is a linear function $\mathbf{H}$ of $\mathbf{y}$. It can be shown that this Gaussian can be re-expressed in terms of $\mathbf{y}$:

$$G_{\mathbf{x}}\left[\mathbf{H}\mathbf{y},\Sigma\right] \propto G_{\mathbf{y}}\left[(\mathbf{H}^T\Sigma^{-1}\mathbf{H})^{-1}\mathbf{H}^T\Sigma^{-1}\mathbf{x},(\mathbf{H}^T\Sigma^{-1}\mathbf{H})^{-1}\right] \tag{5}$$

Proof:

$$\begin{aligned} G_{\mathbf{x}}\left[\mathbf{H}\mathbf{y},\Sigma\right] &= k\exp\left[-0.5\left((\mathbf{x}-\mathbf{H}\mathbf{y})^T\Sigma^{-1}(\mathbf{x}-\mathbf{H}\mathbf{y})\right)\right] \\ &= k\exp\left[-0.5\left(\mathbf{y}^T\mathbf{H}^T\Sigma^{-1}\mathbf{H}\mathbf{y}-\mathbf{y}^T\mathbf{H}^T\Sigma^{-1}\mathbf{x}-\mathbf{x}^T\Sigma^{-1}\mathbf{H}\mathbf{y}^T+\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right)\right] \end{aligned} \tag{6}$$

This exponent is a quadratic function in $\mathbf{y}$ with covariance $(\mathbf{H}^T\Sigma^{-1}\mathbf{H})^{-1}$. Completing the square and absorbing all terms that do not depend on $\mathbf{y}$ into the constant.

$$\begin{aligned} G_{\mathbf{x}}\left[\mathbf{H}\mathbf{y},\Sigma\right] &= k\exp\left[-0.5\left(\mathbf{y}^T\mathbf{H}^T\Sigma^{-1}\mathbf{H}\mathbf{y}-\mathbf{y}^T\mathbf{H}^T\Sigma^{-1}\mathbf{x}-\mathbf{x}^T\Sigma^{-1}\mathbf{H}\mathbf{y}^T+\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right)\right] \\ &= k_2\exp\left[-0.5\left(\mathbf{y}^T\mathbf{H}^T\Sigma^{-1}\mathbf{H}\mathbf{y}-\mathbf{y}^T\mathbf{H}^T\Sigma^{-1}\mathbf{x}-\mathbf{x}^T\Sigma^{-1}\mathbf{H}\mathbf{y}^T\right.\right. \\ &\qquad\qquad \left.\left.+\mathbf{x}^T\Sigma^{-1}\mathbf{H}(\mathbf{H}^T\Sigma^{-1}\mathbf{H})^{-1}\mathbf{H}^T\Sigma^{-1}\mathbf{x}\right)\right] \\ &= k_2\exp\left[-0.5\left((\mathbf{y}-(\mathbf{H}^T\Sigma^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x})^T(\mathbf{H}^T\Sigma^{-1}\mathbf{H})(\mathbf{y}-(\mathbf{H}^T\Sigma^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x})\right)\right] \\ &= k_2G_{\mathbf{y}}\left[(\mathbf{H}^T\Sigma^{-1}\mathbf{H})^{-1}\mathbf{H}^T\Sigma^{-1}\mathbf{x},(\mathbf{H}^T\Sigma^{-1}\mathbf{H})^{-1}\right] \end{aligned} \tag{7}$$

as required.

1.3 Linear Function of a Gaussian

Any linear function of a Gaussian variable is also distributed as a Gaussian. If $\mathbf{x}$ is distributed as $G_x(\mu,\Sigma)$ then the random variable $\mathbf{A}\mathbf{x}+\mathbf{y}$ is also a Gaussian:

$$Pr(\mathbf{A}\mathbf{x}+\mathbf{y}) = G_{\mathbf{x}}\left(\mathbf{A}\mu+\mathbf{y},\mathbf{A}\Sigma\mathbf{A}^T\right)$$

1.4 Convolution of Gaussians

If we convolve two Gaussian distributions with one another, then the result is also a Gaussian so that:

$$G_x(\mathbf{a},\mathbf{A}) * G_x(\mathbf{b},\mathbf{B}) \propto G_x(\mathbf{a}+\mathbf{b},\mathbf{A}+\mathbf{B}).$$

This operation occurs in practice when we draw from one distribution and then draw from a second distribution and sum the results. The distribution of the sum will be equal to the convolution of the two distributions. Note that the result tallies with our earlier proof concerning the sum of n independent Normal variables.

1.5 Matrix Lemma #1

.
Consider the $d \times d$ matrix **P**, the $k \times k$ matrix **R** and the $k \times d$ matrix **H** where **P** and **R** are symmetric, positive definite covariance matrices. The following equality holds:

$$(\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1} = \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1} \tag{8}$$

Proof:

$$\begin{aligned} \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{P}\mathbf{H}^{\mathbf{T}} + \mathbf{H}^T &= \mathbf{H}^T + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{P}\mathbf{H}^T \\ \mathbf{H}^T\mathbf{R}^{-1}(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}) &= (\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\mathbf{P}\mathbf{H}^T \end{aligned} \tag{9}$$

Taking the inverse of both sides:

$$(\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1} = \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1} \tag{10}$$

as required.

1.6 Matrix Lemma #2

.
Consider the $d \times d$ matrix **P**, the $k \times k$ matrix **R** and the $k \times d$ matrix **H** where **P** and **R** are symmetric, positive definite covariance matrices. The following equality holds:

$$(\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1} = \mathbf{P} - \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P} \tag{11}$$

This is known as the *Matrix Inversion Lemma*.
Proof:

$$\begin{aligned} (\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1} &= (\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}(\mathbf{I} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{P} - \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{P}) \\ &= (\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-\mathbf{1}}\mathbf{H})^{-1}\left((\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\mathbf{P} - \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{P}\right) \\ &= \mathbf{P} - (\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}\mathbf{R}^{-1}\mathbf{H}\mathbf{P} \end{aligned} \tag{12}$$

Now, applying Matrix Lemma #1 to the term in brackets:

$$\begin{aligned} (\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1} &= \mathbf{P} - (\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}\mathbf{R}^{-1}\mathbf{H}\mathbf{P} \\ &= \mathbf{P} - \mathbf{P}\mathbf{H}^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P} \end{aligned} \tag{13}$$

as required.

2 Bayes Rule

In this section, we will review Bayes rule, and in particular, we will consider what happens when we have a Gaussian likelihood term, and a Gaussian prior. Bayes' rule is given by:

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$$

In the context of parameter estimation, we have a data term $\mathbf{x}$ and set of parameters θ. Bayes's rule calculates the probability of the parameters given the data so that:

$$Pr(\theta|\mathbf{x}) = \frac{Pr(\mathbf{x}|\theta)Pr(\theta)}{Pr(\mathbf{x})}$$

The term $Pr(\mathbf{x}|\theta)$ is the likelihood of the data given a certain set of parameters, $Pr(\theta)$ is the prior probability of the parameters. The term $Pr(\theta|\mathbf{x})$ is termed the posterior distribution of the parameters.

Example : We are given a single data point, x drawn from a normal distribution with unknown mean μ but known standard deviation σ. We have prior information that the mean μ is itself distributed normally with mean μ_p and standard deviation σ_p. The likelihood and prior terms are hence as follows

$$\begin{array}{rcl} Pr(x|\theta) & = & G_x(\mu, \sigma^2) = G_\mu(x, \sigma^2) \\ Pr(\mu) & = & G_\mu(\mu_p, \sigma_p^2) \end{array}$$

Hence, the posterior for the mean μ can be written

$$Pr(\mu|x) \propto G_\mu(\mu_p, \sigma_p^2) G_\mu(x, \sigma^2)$$

Since these distributions are Gaussian, their product must be Gaussian too. Using the first Gaussian Identity and simplifying, we get:

$$\begin{array}{rcl} Pr(\mu|\mathbf{x}) & \propto & G_\mu[\mu_p, \sigma_p^2] G_\mu[x, \sigma^2] \\ & = & G_\mu\left[\left(\sigma_p^{-2} + \sigma^{-2}\right)^{-1}\left(\sigma_p^{-2}\mu_p + \sigma^{-2}x\right), \left(\sigma_p^{-2} + \sigma^{-2}\right)^{-1}\right] \\ & = & G_\mu\left[\left(\frac{\sigma_p^2\sigma^2}{\sigma_p^2 + \sigma^2}\right)\left(\sigma_p^{-2}\mu_p + \sigma^{-2}x\right), \left(\frac{\sigma_p^2\sigma^2}{\sigma_p^2 + \sigma^2}\right)\right] \\ & = & G_\mu\left[\left(\frac{\sigma^2\mu_p + \sigma_p^2 x}{\sigma_p^2 + \sigma^2}\right), \left(\frac{\sigma_p^2\sigma^2}{\sigma_p^2 + \sigma^2}\right)\right] \end{array} \tag{14}$$

It is worthwhile considering the behavior of the mean of this Gaussian. It is a weighted sum of the term due to the data (x) and the term due to the prior alone (μ_p).

- When the standard deviation of the data is very large compared to the standard deviation of the prior, the prior mean is weighted much more than the mean due to the data.

- When the standard deviation of the data is very small compared to the standard deviation of the prior, the prior mean is weighted much less than the mean due to the data.

In addition, it should be noticed that the variance of the final estimate is less than either the variance due to the data, or the variance due to the prior.

2.1 Bayes Rule For Multi-Dimensional Gaussians

A similar argument to that above leads to the solution for the case where the likelihood and prior are multivariate Gaussian distributions with parameters:

$$\begin{aligned} Pr(\mathbf{x}|\theta) &= G_x(\mu,\Sigma) = G_\mu(\mathbf{x},\Sigma) \\ Pr(\mu) &= G_\mu(\mu_p,\Sigma_p) \end{aligned}$$

The posterior can be shown to be:

$$\begin{aligned} Pr(\mu|\mathbf{x}) &\propto G_\mu[\mu_p,\Sigma_p]G_\mu[x,\Sigma] \\ &= G_\mu\left[\left(\Sigma_p^{-1}+\Sigma^{-1}\right)^{-1}\left(\Sigma_p^{-1}\mu_p+\Sigma^{-1}\mathbf{x}\right),\left(\Sigma_p^{-1}+\Sigma^{-1}\right)^{-1}\right] \end{aligned} \tag{15}$$

The solution has a similar interpretation to before. In directions in space where the prior has less variance than the data, the prior is favored. In directions where the opposite is true, the term due to the data is favored. Of course all of this pre-supposes that each matrix has an inverse. If this isn't the case, this means that the prior (or likelihood) is completely uninformative about one direction in space. This might occur if the measurement is missing data.

3 Temporal Models

3.1 Overview

Consider an evolving system, which is represented by an unknown vector, **x**, which we term the *state*. A concrete example of a possible state vector in computer vision might be the two dimensional position of a tracked object in an image. At each time, $\mathbf{t} = \{1,2\ldots T\}$, we want for form an estimate of $\hat{\mathbf{x}}_t$ of the state, which is in some way optimal. There will be two contributions to this estimate:

1. A set of *measurements*, $\mathbf{z}_t$, which provide information about the state $\mathbf{x}_t$ at time t. We assume that this a generative model: the measurements are derived from the state using a known probability relation, $Pr(\mathbf{z}|\mathbf{x})$.

2. A *time series model*, which says something about the expected way that the system will evolve. In particular, we assume that the state at time t+1 depends only on the state at time t. This is known as the *Markov Assumption*. These states are related by $Pr(\mathbf{x}_{t+1}|\mathbf{x}_t)$.

Let's assume that we have a posterior estimate $Pr(\mathbf{x}_t|\mathbf{z}_1 \ldots \mathbf{z}_t)$ for the state, $\mathbf{x}_t$ at time t given all of the measurements up to that point. Our goal is to calculate a posterior estimate at the time $t+1$, which is written $Pr(\mathbf{x}_{t+1}|\mathbf{z}_1 \ldots \mathbf{z}_{t+1})$. We will do this in two stages. First, we will estimate the new probability distribution of $\mathbf{x}$ at time $t+1$, based only upon the measurements until time t. We term this *state evolution*. This is written as $Pr(\mathbf{x}_{t+1}|\mathbf{z}_1 \ldots \mathbf{z}_t)$. Then we will update to *incorporate the new measurement*. We now examine each of these

3.2 Time Evolution

The aim of the time evolution stage is to move from $p(\mathbf{x}_t|\mathbf{z}_1 \ldots \mathbf{z}_t)$ to $p(\mathbf{x}_{t+1}|\mathbf{z}_1 \ldots \mathbf{z}_t)$. Since no new measurements are observed, this process must be entirely due to the time series model, $Pr(\mathbf{x}_{t+1}|\mathbf{x}_t)$. The update must take account of the uncertainty in the previous estimate of the state:

$$p(\mathbf{x}_{t+1}|\mathbf{z}_1 \ldots \mathbf{z}_t) = \int Pr(\mathbf{x}_{t+1}|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{z}_1 \ldots \mathbf{z}_t) d\mathbf{x}_t \tag{16}$$

This is known as the *Chapman-Kolmogorov* equation. Intuitively, it can be understood as saying: for each possible state $\mathbf{x}_t$ perform the update which will generate a new probability distribution for $\mathbf{x}_{t+1}$, and weight this distribution by the probability of being in the original state $\mathbf{x}_t$. Sum these weighted distributions to yield $Pr(\mathbf{x}_{t+1}|\mathbf{z}_1 \ldots \mathbf{z}_t)$.

3.3 Incorporating a new Measurement

We now consider the result of the time-evolution as a Bayesian Prior, for the full estimate of the state at time t+1. The distribution $p(\mathbf{x}_{t+1}|\mathbf{z}_1 \ldots \mathbf{z}_t)$ represents what we know about the state *before* we consider the measurement $\mathbf{z}_{t+1}$. In order to incorporate the new measurement, we use Bayes' rule.

$$Pr(\mathbf{x}_{t+1}|\mathbf{z}_1 \ldots \mathbf{z}_{t+1}) = \frac{Pr(\mathbf{z}_{t+1}|\mathbf{x}_{t+1}) Pr(\mathbf{x}_{t+1}|\mathbf{z}_1 \ldots \mathbf{z}_t)}{Pr(\mathbf{x}_{t+1})} \tag{17}$$

If all of the relevant distributions are known then the posterior can be calculated exactly. The only assumption that we have made is that the state at time t+1 depends only on the state at time t (the Markov assumption).

3.4 Implementation

In principle, we could represent the two probability distributions $Pr(\mathbf{z}|\mathbf{x})$ and $Pr(\mathbf{x}_{t+1}|\mathbf{x}_t)$ as look up tables. We could also represent the initial state $Pr(\mathbf{x}_0)$ as a non-parametric distribution (look-up table). This should be uniform in the absence of any other information. We could then update the state distribution recursively. Unfortunately, this approach is not practical when the dimension of the state is reasonably large. If we quantize each dimension of the state into just 100 bins then the look-up table for $Pr(\mathbf{x}_{t+1}|\mathbf{x}_t)$ will be of size 100^{2d} where d is the dimensionality of the space. There are similar problems with the distribution $Pr(\mathbf{z}|\mathbf{x})$, and the state distribution itself $Pr(\mathbf{x}_t|\mathbf{z}_1 \ldots \mathbf{z}_t)$. For many applications the state dimension might be quite large: for example, consider tracking an articulated human body model, which might easily have > 25 parameters.

4 The Kalman Filter

4.1 Introduction

The solution to these problems is to use parametric representations for the probability distributions in question. By now, you should have guessed that the distributions in question are going to be Gaussian. Hence, the Chapman-Kolmogorov equation (16) becomes the convolution of a Gaussian with another Gaussian, to yield... a new Gaussian. Similarly, the application of Bayes' rule (17) when we incorporate the new measurement becomes the product of two Gaussians, which yields... a Gaussian. The consequence of this is that instead of maintaining a representation of the whole probability distribution $Pr(\mathbf{x}_t)$, we only have to maintain the mean μ_t and covariance matrix Σ_t.

4.2 Underlying Equations

The probability distributions $Pr(\mathbf{x}_{t+1}|\mathbf{x}_t)$ and $Pr(\mathbf{z}_t|\mathbf{x}_t)$ need to be Gaussian and this constrains the type of temporal updates that we can consider, and also constrains the relationship between the state and the measurements. These two relationships are described by the *state update* and *measurement* equations respectively.

State Update Equation

$$\mathbf{x}_{t+1} = \mathbf{A}_t\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t + \omega_t \tag{18}$$

where $\mathbf{A}_t$ is termed the *state transition matrix* and is square. The terms $\mathbf{B}_t$ and $\mathbf{u}_t$ are known as the *input transmission matrix* and *input* respectively. These are important in control theory, but not for our purposes, and hence will be dropped from now on. Finally ω_t is a noise term which is Gaussian with zero mean and covariance $\mathbf{Q_t}$:

$$\omega_t \sim G_\omega[\mathbf{0}, \mathbf{Q}_t]$$

The state update equation hence involves a linear transformation of the previous state, plus the addition of Gaussian noise. Assuming the estimate of the previous state was Gaussian, the result will also clearly be Gaussian.

Measurement Equation:
The measurement equation describes how the measurement is generated from the state:

$$\mathbf{z}_t = \mathbf{H}_t\mathbf{x}_t + \varepsilon_t \tag{19}$$

The measurements are defined to be a linear function $\mathbf{H}_t$ plus a noise term ε_t which is Gaussian with mean zero and covariance $\mathbf{P}$. i.e.

$$\varepsilon_t \sim G_\varepsilon[\mathbf{0}, \mathbf{P}_t]$$

Note that $\mathbf{H}_t$ is not necessarily square: it is possible to have less measurements than there are state variables. In this case, the matrix $\mathbf{H}$ is not invertible and means that measurements in a single instant of time will not be sufficient to fully determine the state of the system.

5 Kalman Filter Example: 1D Tracking

In order to make these ideas concrete, let's consider tracking a point in 1D over time. The state, x, is now a one-dimensional variable that represents the position of the point.

5.1 State Update Equation

We must define a time series model for the state evolution. Let us assume that on average the point moves with known constant speed, u, but that the speed varies with a Gaussian distribution around this mean:

$$x_{t+1} = x_t + u + \omega \tag{20}$$

where

$$\omega \sim G_\omega[0, \sigma_p^2] \tag{21}$$

The update rule implies that:

$$Pr(x_{t+1}|x_t) = G_x[x_t + u, \sigma_p^2] \tag{22}$$

The state before the update is Gaussian with mean μ_t and standard deviation σ_t:

$$Pr(x_t) = G_x[\mu_t, \sigma_t^2] \tag{23}$$

Now we will apply the Chapman-Kolmogorov relation:

$$\begin{aligned}
Pr(x_{t+1}|z_1 \ldots z_t) &= \int_{-\infty}^{\infty} Pr(x_t) Pr(x_{t+1}|x_t) dx_t \\
&= \int_{-\infty}^{\infty} G_{x_t}[\mu_t, \sigma_t^2] G_{x_{t+1}}[x_t + u, \sigma_p^2] dx_t \\
&= G_{x_{t+1}}[\mu_t + u, \sigma_t^2 + \sigma_p^2] \qquad (24) \\
&= G_{x_{t+1}}[\mu_+, \sigma_+] \qquad (25)
\end{aligned}$$

5.2 Measurement Equation

At each time, we make a noisy measurement of the true position of the point, z, uing the measurement equation:

$$z_t = x_t + \varepsilon \tag{26}$$

where

$$\varepsilon \sim G_\varepsilon[0, \sigma_q^2] \tag{27}$$

so that

$$\begin{aligned}
Pr(z_{t+1}|x_{t+1}) &= G_{z_{t+1}}[x_{t+1}, \sigma_q^2] \\
&= G_{x_{t+1}}[z_{t+1}, \sigma_q^2] \qquad (28)
\end{aligned}$$

Now we use Bayes rule to incorporate the new measurement, using the result of the state update equation as the prior:

$$\begin{aligned} Pr(x_{t+1}|z_1 \ldots z_{t+1}) &= \frac{Pr(z_{t+1}|x_{t+1})Pr(x_{t+1}|z_1 \ldots z_t)}{Pr(x_{t+1})} \\ &= kG_{x_{t+1}}[z_{t+1}, \sigma_q^2]G_{x_{t+1}}[\mu_+, \sigma_+^2] \qquad (29) \\ &= G_{x_{t+1}}\left[\left(\sigma_q^{-2} + \sigma_+^{-2}\right)^{-1}\left(\sigma_q^2 z_{t+1} + \sigma_+^2 \mu_+\right), \left(\sigma_q^{-2} + \sigma_+^{-2}\right)^{-1}\right] \\ & \qquad (30) \end{aligned}$$

Hopefully, this equation should look familiar. Once more, the mean is a weighted sum of the term due to the data, and the term due to the prior. This leads to the following behavior:

- When the time-series model is very strong compared, the prior variance will be relatively small, and the final estimate is closer to the prediction from the times than that from the measurement.
- When the measurement noise is small, the estimate will be predicted largely from the measurement and not rely so heavily on the prior prediction from the time-series.

Note that even if the measurement is entirely missing (infinite variance), the time series model still allows the system to make a prediction.

6 Kalman Filter Example: N-D Tracking

We now extend the results of the previous section to the N-dimensional case. The main extra complication is that the measurement matrix, **H** cannot necessarily be inverted.

6.1 State Update Equation

We must define a time series model for the state evolution. We will consider the general case where the new state is a linear transformation of the previous state.

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \omega \qquad (31)$$

where

$$\omega \sim G_\omega[0, \mathbf{P}] \qquad (32)$$

The update rule implies that:

$$Pr(\mathbf{x}_{t+1}|\mathbf{x}_t) = G_{\mathbf{x}_{t+1}}[\mathbf{A}\mathbf{x}_t, \mathbf{P}] \qquad (33)$$

The state before the update is Gaussian with mean μ_t and variance Σ_t:

$$Pr(\mathbf{x}_t) = G_x[\mu_t, \Sigma_t] \qquad (34)$$

Now we will apply the Chapman-Kolmogorov relation:

$$
\begin{aligned}
Pr(\mathbf{x}_{t+1}|\mathbf{z}_1\ldots\mathbf{z}_t) &= \int_{-\infty}^{\infty} Pr(\mathbf{x}_t)Pr(\mathbf{x}_{t+1}|\mathbf{x}_t)d\mathbf{x}_t \\
&= \int_{-\infty}^{\infty} G_{\mathbf{x}_t}[\mu_t,\Sigma_t]G_{\mathbf{x}_{t+1}}[\mathbf{A}\mathbf{x}_t,\mathbf{P}]d\mathbf{x}_t \\
&= G_{x_{t+1}}[\mathbf{A}\mu_t,\mathbf{A}\Sigma_t\mathbf{A}^T+\mathbf{P}] \qquad (35) \\
&= G_{x_{t+1}}[\mu_+,\Sigma_+] \qquad (36)
\end{aligned}
$$

6.2 Measurement Equation

At each time, we make a noisy measurement of the true position of the point, z, uing the measurement equation:

$$
\mathbf{z}_t = \mathbf{H}\mathbf{x}_t + \varepsilon \tag{37}
$$

where

$$
\varepsilon \sim G_\varepsilon[0,\mathbf{Q}] \tag{38}
$$

so that

$$
\begin{aligned}
Pr(\mathbf{z}_{t+1}|\mathbf{x}_{t+1}) &= G_{\mathbf{z}_{t+1}}[\mathbf{H}\mathbf{x}_{t+1},\mathbf{Q}] \\
&= G_{x_{t+1}}[(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{z},(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H})^{-1}] \qquad (39)
\end{aligned}
$$

Now we use Bayes rule to incorporate the new measurement, using the result of the state update equation as the prior:

$$
\begin{aligned}
Pr(\mathbf{x}_{t+1}|\mathbf{z}_1\ldots\mathbf{z}_{t+1}) &= \frac{Pr(\mathbf{z}_{t+1}|\mathbf{x}_{t+1})Pr(\mathbf{x}_{t+1}|\mathbf{z}_1\ldots\mathbf{z}_t)}{Pr(\mathbf{x}_{t+1})} \\
&= kG_{x_{t+1}}[(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{z},(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H})^{-1}]G_{x_{t+1}}[\mu_+,\Sigma_+] \\
&= G_{x_{t+1}}[(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{z},(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H})^{-1}]G_{x_{t+1}}[\mu_+,\Sigma_+] \\
&= G_{x_{t+1}}\left[\left(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H}+\Sigma_+^{-1}\right)^{-1}\left(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{z}+\Sigma_+^{-1}\mu_+\right),(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H}+\Sigma_+^{-1})^{-1}\right] \\
&\qquad (40)
\end{aligned}
$$

This is a perfectly acceptable answer, but for various reasons, including numerical concerns, the result is usually presented in a slightly different form. We define the *Kalman Gain* as:

$$
\mathbf{K} = \Sigma_+\mathbf{H}^T(\mathbf{Q}+\mathbf{H}\Sigma_+\mathbf{H}^T)^{-1} \tag{41}
$$

We will first consider the mean term. Substituting in the Kalman Gain and using the Matrix Lemmas #1 and #2.

$$
\begin{aligned}
\left(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H}+\Sigma_+^{-1}\right)^{-1}\left(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{z}+\Sigma_+^{-1}\mu_+\right)& \\
&= \mathbf{K}\mathbf{z}+\left(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H}+\Sigma_+^{-1}\right)^{-1}\Sigma_+^{-1}\mu_+ \\
&= \mathbf{K}\mathbf{z}+\left(\Sigma_+-\Sigma_+\mathbf{H}^T(\mathbf{H}\Sigma_+\mathbf{H}^T+\mathbf{Q})^{-1}\mathbf{H}\Sigma_+\right)\Sigma_+^{-1}\mu_+ \\
&= \mathbf{K}\mathbf{z}+\mu_+-\Sigma_+\mathbf{H}^T(\mathbf{H}\Sigma_+\mathbf{H}^T+\mathbf{Q})^{-1}\mathbf{H}\mu_+ \\
&= \mathbf{K}\mathbf{z}+\mu_+-\mathbf{K}\mathbf{H}\mu_+ \\
&= \mu_++\mathbf{K}(\mathbf{z}-\mathbf{H}\mu_+) \qquad (42)
\end{aligned}
$$

The expression in brackets is known as the *innovation*, and is the difference between the actual measurements, $\mathbf{z}_{t+1}$ and the predicted measurements based on the prior estimate of the state. It is easy to see why $\mathbf{K}$ is termed the Kalman gain: this determines the amount that the measurements contribute to the new estimate. If the Kalman gain is small then this implies that the variance on the measurements is great compared to the prior. If the Kalman gain is large then this suggests that the measurements are more reliable than the prior and should be weighted more highly.
Now let’s consider the covariance term, and again using Matrix identity #1.

$$
\begin{aligned}
(\mathbf{H}^T\mathbf{Q}^{-1}\mathbf{H}+\Sigma_+^{-1})^{-1} &= \Sigma_+-\Sigma_+\mathbf{H}^T(\mathbf{H}\Sigma_+\mathbf{H}+\mathbf{Q})^{-1}\mathbf{H}\Sigma_+ \\
&= \Sigma_+-\mathbf{K}\mathbf{H}\Sigma_+ \\
&= (\mathbf{I}-\mathbf{K}\mathbf{H})\Sigma_+ \qquad (43)
\end{aligned}
$$

So, the final result is:

$$
Pr(\mathbf{x}_{t+1}|\mathbf{z}_1\ldots\mathbf{z}_{t+1}) = G_{\mathbf{x}_{t+1}}[\mu_++\mathbf{K}(\mathbf{z}-\mathbf{H}\mu_+),(\mathbf{I}-\mathbf{K}\mathbf{H})\Sigma_+] \qquad (44)
$$

7 Kalman Filter Equations In Full

In the previous section, we have derived the Kalman filter equations from a Bayesian perspective. There are several other ways to arrive at the same conclusion, and the results are usually expressed in terms of update rules. We now summarize the aim, assumptions and update rule for the Kalman filter:
The aim of the Kalman filter is to calculate the posterior probability $Pr(\mathbf{x}_{t+1}|\mathbf{z}_1\ldots z_{t+1})$ state at time t+1 given the probability $Pr(\mathbf{x}_t|\mathbf{z}_1\ldots z_t)$ state at time t and a measurement $\mathbf{z}_{t+1}$ at time t+1. We assume the following dynamical model:

$$
\begin{aligned}
\mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t+\omega \qquad (45)\\
\mathbf{z}_t &= \mathbf{H}\mathbf{x}_t+\varepsilon \qquad (46)
\end{aligned}
$$

where

$$
\begin{aligned}
\varepsilon &\sim G_\varepsilon[0,\mathbf{Q}] \qquad (47)\\
\omega &\sim G_\omega[0,\mathbf{P}] \qquad (48)
\end{aligned}
$$

We make the following assumptions:

$$\begin{aligned}
E[\mathbf{x}_t, \varepsilon_k] &= 0 && \forall t, k \\
E[\mathbf{x}_t, \omega_k] &= 0 && t \leq k \\
E[\mathbf{z}_t, \varepsilon_k] &= 0 && t \leq k-1 \\
E[\mathbf{z}_t, \omega_k] &= 0 && t \leq k \\
E[\varepsilon_t, \omega_k] &= 0 && \forall t, k \\
E[\varepsilon_t, \varepsilon_t] &= 0 && t \neq k \\
E[\omega_t, \omega_k] &= 0 && t \neq k
\end{aligned}$$

The Kalman filter update equations are then given by:

$$\begin{aligned}
\text{State Prediction:} \qquad \mu_+ &= \mathbf{A}\mu_t \\
\text{Covariance Prediction:} \qquad \Sigma_+ &= \mathbf{A}\Sigma_t\mathbf{A}^T + \mathbf{P} \\
\text{State Update:} \qquad \mu_{t+1} &= \mu_+ + \mathbf{K}(\mathbf{z}_{t+1} - \mathbf{H}\mu_+) \\
\text{Covariance Update:} \qquad \Sigma_{t+1} &= (\mathbf{I} - \mathbf{K}\mathbf{H})\Sigma_+
\end{aligned} \tag{49}$$

where

$$\mathbf{K} = \Sigma_+\mathbf{H}^T(\mathbf{Q} + \mathbf{H}\Sigma_+\mathbf{H}^T)^{-1} \tag{50}$$

7.1 Time Invariance

In the above derivation, we have assumed that the measurement matrix, state update matrix, and the covariance matrices are constant at each time step. However, it may be that each of these changes on every time step depending on the application. In this case, we carry on as normal using the time varying versions of **A**, **H**, **P** and **Q**.

7.2 Learning Parameters

It is possible to learn the optimal state update matrix **A** and measurement matrix **H** if there are not a-priori grounds for setting them. The basic principle is to maximize the overall likelihood of the data. This is a chicken and egg type problem: as soon as we change either **H** or **A** the state estimates $Pr(\mathbf{x}_t)$ will also change. The standard approach to this kind of problem is to use the Expectation-Maximization algorithm (E-M algorithm). The idea is to first calculate the optimal states with fixed **A** and **H** and then to fix the state estimates and try to improve the current **A** and **H**.

7.3 Including Velocity

In the 1D example, we introduced a rather artifical case where we knew the velocity of the point. It is possible to estimate velocity (and indeed acceleration and higher derivatives) by building them into the time update and measurement equations. Consider tracking a point in 2D. We augment the state vector so that

$$\mathbf{x}_t = \begin{bmatrix} \dot{x}_t \\ \dot{y}_t \\ x_t \\ y_t \end{bmatrix} \tag{51}$$

where x and y represent the position of the point in 2D and $\dot{x}$ and $\dot{y}$ represent the (unknown velocity). We now introduce the augmented state update equation and measurement equations:

$$\begin{bmatrix} \dot{x}_{t+1} \\ \dot{y}_{t+1} \\ x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{x}_t \\ \dot{y}_t \\ x_t \\ y_t \end{bmatrix} + \varepsilon \tag{52}$$

$$\mathbf{z}_{t+1} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{x}_{t+1} \\ \dot{y}_{t+1} \\ x_{t+1} \\ y_{t+1} \end{bmatrix} + \omega \tag{53}$$

We then proceed as usual, substituting these new matrices for **A** and **H**.

8 Extended Kalman Filter

The extended Kalman filter deals with the case where the time-update and/or the measurement equations are arbritrary non-linear functions of the form:

$$\mathbf{x}_{t+1} \quad = \quad a(\mathbf{x}_t, \varepsilon_t) \tag{54}$$

$$\mathbf{z}_t = \quad = \quad h(\mathbf{x}_t, \omega_t) \tag{55}$$

In practice, we do not know the values of the noise ε_t and ω_t. By linearizing the equations, we can generate rules for updating the mean and co-variance as before:

$$\begin{array}{rrl} \text{State Prediction:} & \mu_+ = & a(\mu_t, \mathbf{0}) \\ \text{Covariance Prediction:} & \Sigma_+ = & \mathbf{A}\Sigma_t\mathbf{A}^T + \mathbf{E}\mathbf{P}\mathbf{E}^T \\ \text{State Update:} & \mu_{t+1} = & \mu_+ + \mathbf{K}(\mathbf{z}_{t+1} - h(\mu_+, 0)) \\ \text{Covariance Update:} & \Sigma_{t+1} = & (\mathbf{I} - \mathbf{K}\mathbf{H})\Sigma_+ \end{array} \tag{56}$$

where:

$$\mathbf{K} = \Sigma_+\mathbf{H}^T(\mathbf{W}\mathbf{Q}\mathbf{W}^T + \mathbf{H}\Sigma_+\mathbf{H}^T)^{-1} \tag{57}$$

and $\mathbf{A}, \mathbf{H}, \mathbf{E}$ and $\mathbf{W}$ are the Jacobian matrices:

$$\begin{aligned}\mathbf{A}_{ij} &= \frac{\partial a_i(\mu_t, \mathbf{0})}{dx_j} &(58)\\ \mathbf{E}_{ij} &= \frac{\partial a_i(\mu_t, \mathbf{0})}{d\varepsilon_j} &(59)\\ \mathbf{H}_{ij} &= \frac{\partial h_i(\mu_+, \mathbf{0})}{dx_j} &(60)\\ \mathbf{W}_{ij} &= \frac{\partial h_i(\mu_+, \mathbf{0})}{d\omega_j} &(61)\\ & &(62)\end{aligned}$$

9 Smoothing

The estimates that we have provided so far depend only on measurements up to the current point in time. In some circumstances however, we want to calculate an optimal estimate of the state based on the entire time-series, including those that came after the current time, t. This refinement of the on-line estimates is known as *smoothing*. There are several possible schemes, including:

- *Fixed Point Smoother:* Here we wish to estimate the state $\mathbf{x}_\tau$ at a fixed point in time as time evolves. i.e. we wish to calculate $Pr(\mathbf{x}_\tau|\mathbf{z}_1 \ldots \mathbf{z}_t)$ for all values of t.
- *Fixed Lag Smoother:* This is an on-line scheme in which the optimal estimate for a state at time $t-\tau$ is calculated based on measurements up to time t, where τ is the time lag. i.e. we wish to calculate $Pr(\mathbf{x}_{t-\tau}|\mathbf{z}_1 \ldots \mathbf{z}_t)$.
- *Fixed Interval Smoother:* We have a fixed time interval of measurements and want to calculate the optimal state estimate based on all of these measurements. In other words, instead of calculating $Pr(\mathbf{x}_t|\mathbf{z}_1 \ldots \mathbf{z}_t)$ we now estimate $Pr(\mathbf{x}_t|\mathbf{z}_1 \ldots \mathbf{z}_T$ where T is the total length of the interval.

9.1 Fixed Point Smoother

We wish to calculate $Pr(\mathbf{x}_\tau|\mathbf{z}_1 \ldots \mathbf{z}_t)$ for all values of $t > \tau$. There is a simple and elegant way to accomplish this. Let's assume that we have calculated $Pr(\mathbf{x}_\tau|\mathbf{z}_1 \ldots \mathbf{z}_\tau)$. Now, we augment the state evolution and measurement equations as follows:

$$\begin{bmatrix}\mathbf{x}_{t+1}\\ \mathbf{x}^\tau_{t+1}\end{bmatrix} = \begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{x}_t\\ \mathbf{x}^\tau_t\end{bmatrix} + \begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix}\omega \tag{63}$$

$$\mathbf{z}_t = \begin{bmatrix}\mathbf{H} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}_t\\ \mathbf{x}^\tau_t\end{bmatrix} + \varepsilon \tag{64}$$

where $\mathbf{x}^\tau_t$ is the state at time τ given t measurements. Notice that these augmented equations have the same general form as the original time update and measurement equations. To perform fixed lag smoothing, we hence define the new matrices:

$$
\begin{aligned}
\mathbf{x}'_t &= \begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{x}^{\tau}_{t+1} \end{bmatrix} \\
\mathbf{A}' &= \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \\
\mathbf{H}' &= \begin{bmatrix} \mathbf{H} & \mathbf{0} \end{bmatrix}
\end{aligned}
$$

and then proceed as normal with these matrices in place of the original ones.

9.2 Fixed Lag Smoother

This is an on-line scheme in which the optimal estimate for a state at time $t-\tau$ is calculated based on measurements up to time t, where τ is the time lag. i.e. we wish to calculate $Pr(\mathbf{x}_{t-\tau}|\mathbf{z}_1 \ldots \mathbf{z}_t)$. The principle of calculating the fixed-lag solution is similar to the fixed point solution. We calculate an augmented set of equations:

$$
\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{x}_{(1)t+1} \\ \mathbf{x}_{(2)t+1} \\ \vdots \\ \mathbf{x}_{(\tau)t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{t} \\ \mathbf{x}_{(1)t} \\ \mathbf{x}_{(2)t} \\ \vdots \\ \mathbf{x}_{(\tau)t} \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \omega \tag{65}
$$

where $\mathbf{x}_{(l)t}$ denotes the state at time t with a measurement lag of l. The measurement equation becomes:

$$
\mathbf{z}_t = \begin{bmatrix} \mathbf{H} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{t} \\ \mathbf{x}_{(1)t} \\ \mathbf{x}_{(2)t} \\ \vdots \\ \mathbf{x}_{(\tau)t} \end{bmatrix} + \varepsilon_t \tag{66}
$$

9.3 Fixed Interval Smoother

We now introduce the notation $\mu_{t|t}$ to represent the estimate of the mean at time t given t measurements. For a complete times series of length M we wish to calculate the estimates of the mean and variance given the whole sequence, $\mu_{t|M}$ and $\Sigma_{t|M}$. We assume that we have already calculated the forward sequence. Then we perform the following recursions back from the end to the beginning.

$$
\mu_{t-1|M} = \mu_{t-1|t-1} + \Sigma_{t-1|t-1}\mathbf{A}^T_{t-1}\Sigma^{-1}_{t|t-1}(\mu_{t|M} - \mu_{t|t-1}) \tag{67}
$$

$$
\Sigma_{t-1|M} = \Sigma_{t-1|t-1} + \mathbf{C}_{t-1}(\Sigma_{t|M} - \Sigma_{t|t-1})\mathbf{C}^T_{t-1} \tag{68}
$$

where

$$
\mathbf{C}_{t-1} = \Sigma_{t-1|t-1}\mathbf{A}^T_{t-1}\Sigma^{-1}_{t|t-1} \tag{69}
$$