A Note on Kalman Filter

Tsang-Kai Chang
Department of Electrical and Computer Engineering
University of California, Los Angeles
Los Angeles, CA 90025
tsangkaichang@ucla.edu

November 12, 2020

1 Introduction

"Filtering is the science of finding the law of a process given a partial observation of it." (1) This is a very common challenge in a lot of (engineering) systems, especially robotic systems. Often time, we do not have access to the state of the process we are interested in; instead, we estimate the state with all the partial observations we have. Among all filtering algorithms, this note focuses on the simplest one: the Kalman filter. Developed around 1960s, Kalman filter is not only theoretically solid but is also widely applicable. In its early development, Kalman filter was incorporated in the Apollo navigation computer for the nonlinear trajectory estimation problem.

The goal of this note is to provide an introductory tutorial to this method, and establish the foundation for more advanced material. Even though there are various approaches to consider the Kalman filter, I take a probabilistic point of view to set it up. Personally, I find this approach most elegant and extendable. Typical Kalman filter tutorials are created every once in a while, for example (2).

The notation in this note follows this convention: Random vectors are in boldface. Capital symbols are reserved for matrices.

2 Multivariate Gaussian Distributions

A random vector $\mathbf{x} \in \mathbb{R}^n$ is multivariate Gaussian if its pdf is given by

$$p_{\mathbf{x}}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)\right).$$

The vector $\mu \in \mathbb{R}^n$ is called the mean, and the positive definite matrix $\Sigma \in \mathbb{R}^{n\times n}$ is called the covariance matrix. If $\mathbf{x}$ has the distribution $p_{\mathbf{x}}(x; \mu, \Sigma)$, we can simply express as $\mathbf{x} \sim N(\mu, \Sigma)$.

The quantity $\sqrt{(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)}$ is known as the Mahalanobis distance. The notation $\|x\|_M^2 = x^\mathsf{T} M^{-1} x$ is used for the squared Mahalanobis distance with covariance matrix M. Thus, the PDF is also expressed as

$$p_{\mathbf{x}}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\|x-\mu\|_\Sigma^2\right).$$

This squared term makes the contour of the multivariate Gaussian distribution elliptical.

Proposition 1 (Affine transformation)**.** *If $\mathbf{x}$ is multivariate Gaussian with distribution $N(\mu, \Sigma)$, then $A\mathbf{x} + b$ has distribution $N(A\mu + b, A\Sigma A^\mathsf{T})$.*

Proposition 2 (Independence)**.** *Given a Gaussian random vector $[\mathbf{x}^\mathsf{T}, \mathbf{y}^\mathsf{T}]^\mathsf{T}$, $\mathbf{x}$ and $\mathbf{y}$ are independent if and only if the covariance matrix is block diagonal.*

Gaussian random vectors have a lot of interesting properties. For our purpose to understand the Kalman filter, we focus on the following two operations.

Proposition 3 (Addition). *If $\mathbf{x} \sim N(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$ and $\mathbf{y} \sim N(\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}})$ are two independent multivariate Gaussian random vectors with the same dimension, then $\mathbf{x}+\mathbf{y} \sim N(\mu_{\mathbf{x}}+\mu_{\mathbf{y}}, \Sigma_{\mathbf{x}}+\Sigma_{\mathbf{y}})$.*

By obtaining the characteristic functions of both random vectors, the result is immediate. Or you can think this as the result from the aforementioned two properties.

Proposition 4 (Conditioning). *If $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ are jointly Gaussian*

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_{\mathbf{x}}\\ \mu_{\mathbf{y}}\end{bmatrix}, \begin{bmatrix}\Sigma_{\mathbf{x}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{y}}\end{bmatrix}\right),$$

then the conditional distribution of $\mathbf{x}$ given $\mathbf{y} = y$ is still Gaussian with distribution $N\left(\mu_{\mathbf{x}} + \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}(y-\mu_{\mathbf{y}}), \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\Sigma_{\mathbf{yx}}\right)$.

Starting with the undergraduate definition of conditional distribution

$$p_{\mathbf{x}}(x|\mathbf{y}=y) = \frac{p(x_1, x_2)}{\int_{x\in\mathbb{R}^m} p(x_1,x_2)dx}, \tag{1}$$

the result can be obtained after some lengthy calculation.

Even though obtaining the distribution is sufficient, I layout a stronger argument based on conditional expectation. By doing so, we will have a consistent argument even in the continuous-time case, or the Kalman-Bucy filter.

Proposition 5 (Conditional expectation). *If $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ are jointly Gaussian*

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_{\mathbf{x}}\\ \mu_{\mathbf{y}}\end{bmatrix}, \begin{bmatrix}\Sigma_{\mathbf{x}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{y}}\end{bmatrix}\right),$$

then

$$\mathsf{E}[\mathbf{x}|\mathbf{y}] = \mu_{\mathbf{x}} + \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}(\mathbf{y}-\mu_{\mathbf{y}}). \tag{2}$$

Proof. Let's begin with a simpler case that $\mathbf{x}$ and $\mathbf{y}$ are zero-mean, or

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\Sigma_{\mathbf{x}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{y}}\end{bmatrix}\right).$$

We define $\mathbf{x} = \hat{\mathbf{x}} + \tilde{\mathbf{x}}$ with

$$\begin{aligned}\hat{\mathbf{x}} &= \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\mathbf{y},\\ \tilde{\mathbf{x}} &= \mathbf{x} - \hat{\mathbf{x}}\\ &= \mathbf{x} - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\mathbf{y}.\end{aligned}$$

Since

$$\begin{bmatrix}\tilde{\mathbf{x}}\\ \mathbf{y}\end{bmatrix} = \begin{bmatrix}I_m & -\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\\ 0 & I_n\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix},$$

the random vector $[\tilde{\mathbf{x}}^\mathsf{T}, \mathbf{y}^\mathsf{T}]^\mathsf{T}$ is still Gaussian. Furthermore, it is zero-mean with covariance

$$\begin{aligned}\mathsf{Cov}\left(\begin{bmatrix}\tilde{\mathbf{x}}\\ \mathbf{y}\end{bmatrix}\right) &= \begin{bmatrix}I_m & -\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\\ 0 & I_n\end{bmatrix}\begin{bmatrix}\Sigma_{\mathbf{x}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{y}}\end{bmatrix}\begin{bmatrix}I_m & -\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\\ 0 & I_n\end{bmatrix}^\mathsf{T}\\ &= \begin{bmatrix}\Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\Sigma_{\mathbf{yx}} & 0\\ 0 & \Sigma_{\mathbf{y}}\end{bmatrix}.\end{aligned}$$

Therefore, $\tilde{\mathbf{x}}$ and $\mathbf{y}$ are independent. We now have

$$\begin{aligned}\mathsf{E}[\mathbf{x}|\mathbf{y}] &= \mathsf{E}[\hat{\mathbf{x}} + \tilde{\mathbf{x}}|\mathbf{y}]\\ &= \mathsf{E}[\hat{\mathbf{x}}|\mathbf{y}] + \mathsf{E}[\tilde{\mathbf{x}}|\mathbf{y}]\\ &= \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\mathbf{y} + \mathsf{E}[\tilde{\mathbf{x}}]\\ &= \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\mathbf{y}.\end{aligned}$$

For the general case that

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_{\mathbf{y}} \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathbf{x}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{y}} \end{bmatrix} \right),$$

we have to find $\mathbf{x} = \hat{\mathbf{x}} + \tilde{\mathbf{x}}$ such that $\tilde{\mathbf{x}}$ is still zero-mean and is independent of $\mathbf{y}$. By choosing

$$\begin{aligned} \hat{\mathbf{x}} &= \mu_{\mathbf{x}} + \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}(\mathbf{y} - \mu_{\mathbf{y}}), \\ \tilde{\mathbf{x}} &= \mathbf{x} - \hat{\mathbf{x}} \\ &= \mathbf{x} - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\mathbf{y}, \end{aligned}$$

the same argument holds that $\tilde{\mathbf{x}}$ and $\mathbf{y}$ are independent, and consequently the main proposition follows. □

The above proof actually decomposes $\mathbf{x}$ into dependent and independent part. The conditional PDF of $\mathbf{x}$ can then be easily determined. Given a value of $\mathbf{y}$, the random vector $\hat{\mathbf{x}}$ becomes a known constant, but the distribution of $\tilde{\mathbf{x}}$ is unaffected, since $\tilde{\mathbf{x}}$ is independent of $\mathbf{y}$. Therefore, the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ is the same as the unconditional distribution of $\tilde{\mathbf{x}}$, shifted by $\hat{\mathbf{x}}$. From the proof, $\mathsf{Cov}(\mathbf{x}|\mathbf{y}) = \mathsf{Cov}(\tilde{\mathbf{x}}) = \Sigma_{\mathbf{x}} - \Sigma_{\mathbf{xy}}\Sigma_{\mathbf{y}}^{-1}\Sigma_{\mathbf{yx}}$.

This part of the text is modified and expanded from (3).

3 Kalman Filter

We now consider a linear dynamic model with time dynamic model:

$$\mathbf{s}_{t+1} = F\mathbf{s}_t + \mathbf{w}_t \in \mathbb{R}^n, \quad t = 0, 1, \ldots, \tag{3}$$

and the observation model:

$$\mathbf{o}_t = H\mathbf{s}_t + \mathbf{v}_t \in \mathbb{R}^m, \quad t = 0, 1, \ldots. \tag{4}$$

We assume that $\mathbf{s}_0, \mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{v}_0, \mathbf{v}_1, \ldots,$ are jointly Gaussian and independent. Furthermore, $\mathsf{E}[\mathbf{w}_t] = 0$ and $\mathsf{E}[\mathbf{w}_t\mathbf{w}_t^{\mathsf{T}}] = Q$. $\mathsf{E}[\mathbf{v}_t] = 0$ and $\mathsf{E}[\mathbf{v}_t\mathbf{v}_t^{\mathsf{T}}] = R$.

3.1 Theory and Derivation

While we do not have access to $\mathbf{s}_t$ directly, we wish to estimate $\mathbf{s}_t$ optimally with the observations. The following theorem tells that what we are looking for is nothing but conditional expectation.

Theorem 1. *The σ-algebra generated by $\mathbf{z}$ is denoted by $\sigma(\mathbf{z})$, which contains all Borel-measurable functions of $\mathbf{z}$. Then*

$$\arg\min_{\mathbf{y}\in\sigma(\mathbf{z})} \mathsf{E}[||\mathbf{x} - \mathbf{y}||^2] = \mathsf{E}[\mathbf{x}|\mathbf{y}]. \tag{5}$$

In fact, it is shown in (4) that the conditional expectation is not only the optimal estimator for the least-mean-square error, but is also optimal for all Bregman loss functions.

According to Theorem 1, the optimal estimator for $\mathbf{s}_{t+1}$ given $\mathbf{o}_{1:t} = \{\mathbf{o}_1, \ldots, \mathbf{o}_t\}$, is just $\mathsf{E}[\mathbf{s}_{t+1}|\mathbf{o}_{1:t}]$. Similarly, the optimal estimator for $\mathbf{s}_{t+1}$ given $\mathbf{o}_{1:t+1}$ is $\mathsf{E}[\mathbf{s}_{t+1}|\mathbf{o}_{1:t+1}]$. In reality, observations are no longer random variable but sampled values. The optimal estimators $\mathsf{E}[\mathbf{s}_{t+1}|\mathbf{o}_{1:t} = o_{1:t}]$ and $\mathsf{E}[\mathbf{s}_{t+1}|\mathbf{o}_{1:t+1} = o_{1:t+1}]$ are just deterministic functions of $o_{1:t}$ and $o_{1:t+1}$.

The Kalman filter is actually an efficient algorithm to calculate the conditional distributions of $\mathbf{s}_{t+1}$ given $\mathbf{o}_{1:t} = o_{1:t}$ and $\mathbf{o}_{1:t+1} = o_{1:t+1}$, recursively.

To simply the notation, we define $\mathbf{s}_{t|\tau}$ to denote the conditional distribution of $\mathbf{s}_t$ given $\mathbf{s}_0, \mathbf{o}_{1:\tau}$. The condition on $\mathbf{s}_0$ is to ensure that the entire HMM is well-define. We won't explicitly mention it in the following derivation.

We begin to derive the time propagation update for the Kalman filter. Based on the time dynamic model (2), we have

$$\mathbf{s}_{t+1|t} = F\mathbf{s}_{t|t} + \mathbf{w}_t.$$

If $\mathbf{s}_{t|t}$ follows the distribution $N(\bar{s}_{t|t}, \Sigma_{t|t})$, it is obvious that $\mathbf{s}_{t+1|t}$ is also Gaussian with mean and covariance

$$\bar{s}_{t+1|t} = F\bar{s}_{t|t}, \tag{6}$$

$$\Sigma_{t+1|t} = F\Sigma_{t|t}F^{\mathsf{T}} + Q. \tag{7}$$

For the observation update, we begin with the observation model (3) and condition each random variable on the information up to now:

$$\begin{aligned}\mathbf{o}_{t+1}|\mathbf{o}_{1:t} &= H\mathbf{s}_{t+1}|\mathbf{o}_{1:t} + \mathbf{v}_{t+1}|\mathbf{o}_{1:t}\\ &= H\mathbf{s}_{t+1}|\mathbf{o}_{1:t} + \mathbf{v}_{t+1}.\end{aligned}$$

We can see that

$$\begin{bmatrix}\mathbf{s}_{t+1}|\mathbf{o}_{1:t}\\ \mathbf{o}_{t+1}|\mathbf{o}_{1:t}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\bar{s}_{t+1|t}\\ H\bar{s}_{t+1|t}\end{bmatrix}, \begin{bmatrix}\Sigma_{t+1|t} & \Sigma_{t+1|t}H^{\mathsf{T}}\\ H\Sigma_{t+1|t} & H\Sigma_{t+1|t}H^{\mathsf{T}} + R\end{bmatrix}\right).$$

Now, $\mathbf{s}_{t+1|t+1} = (\mathbf{s}_{t+1}|\mathbf{o}_{1:t})|(\mathbf{o}_{t+1}|\mathbf{o}_{1:t})$. With the formula that we derived in the previous section and with $\mathbf{o}_{1:t+1} = o_{1:t+1}$,

$$\bar{s}_{t+1|t+1} = \bar{s}_{t+1|t} + \Sigma_{t+1|t}H^{\mathsf{T}}\left(H\Sigma_{t+1|t}H^{\mathsf{T}} + R\right)^{-1}\left(o_{t+1} - H\bar{s}_{t+1|1:t}\right), \tag{8}$$

$$\Sigma_{t+1|t+1} = \Sigma_{t+1|t} - \Sigma_{t+1|t}H^{\mathsf{T}}\left(H\Sigma_{t+1|t}H^{\mathsf{T}} + R\right)^{-1}H\Sigma_{t+1|t}. \tag{9}$$

Most of the textbooks might write (7) and (8) in

$$\begin{aligned}\bar{s}_{t+1|t+1} &= \bar{s}_{t+1|t} + K_{t+1}(o_{t+1} - H\bar{s}_{t+1|t}),\\ \Sigma_{t+1|t+1} &= (I - K_{t+1}H)\Sigma_{t+1|t},\end{aligned}$$

with $K_{t+1} = \Sigma_{t+1|t}H^{\mathsf{T}}\left(H\Sigma_{t+1|t}H^{\mathsf{T}} + R\right)^{-1}$, and they call K_t the Kalman gain.

3.2 Systems with Constant Inputs

We can easily extend the result of Kalman filter to incorporate constant inputs. For example, we consider a more general system model

$$\mathbf{s}_{t+1} = F\mathbf{s}_t + Gu_t + \mathbf{w}_t, \quad t = 0, 1, \ldots, \tag{10}$$

and

$$\mathbf{o}_t = H\mathbf{s}_t + J\lambda_t + \mathbf{v}_t, \quad t = 0, 1, \ldots. \tag{11}$$

u_t and λ are not random vectors, but are parameters in the system models.

Following the identical derivation, the update equations of the Kalman filter become

$$\bar{s}_{t+1|t} = F\bar{s}_{t|t} + Gu_t, \tag{12}$$

$$\Sigma_{t+1|t} = F\Sigma_{t|t}F^{\mathsf{T}} + Q. \tag{13}$$

and

$$\bar{s}_{t+1|t+1} = \bar{s}_{t+1|t} + \Sigma_{t+1|t}H^{\mathsf{T}}\left(H\Sigma_{t+1|t}H^{\mathsf{T}} + R\right)^{-1}\left(o_{t+1} - H\bar{s}_{t+1|1:t} - J\lambda_t\right), \tag{14}$$

$$\Sigma_{t+1|t+1} = \Sigma_{t+1|t} - \Sigma_{t+1|t}H^{\mathsf{T}}\left(H\Sigma_{t+1|t}H^{\mathsf{T}} + R\right)^{-1}H\Sigma_{t+1|t}. \tag{15}$$

We can see that in the presence of constant inputs, only the means of the estimation distribution change, but the covariance matrices remain the same.

3.3 Convergence

In the Kalman filtering setting, the covariance matrix represents the estimation uncertainty. Whether the estimation uncertainty stays small, or at least bounded, becomes very important to the estimation performance. We now discuss when the covariance matrix of a Kalman filter will converge.

First, we rewrite (8) as

$$\Sigma^{-1}_{t+1|t+1} = \Sigma^{-1}_{t+1|t} + H^{\mathsf{T}}R^{-1}H. \tag{16}$$

by the matrix inversion lemma. In other words, the covariance evolving equations of the time update and the observation update are dual. While we can say that the time update adds a positive definite matrix on the covariance matrix, the observation update adds a positive definite matrix on the inverse of the covariance matrix. Therefore, we can roughly say that the time update increases the uncertainty while the observation update decrease the uncertainty.

Now, by combining (6) and (8), we have

$$\Sigma_{t+1|t} = F(\Sigma_{t|t-1}^{-1} + H^\top R^{-1} H)^{-1} F^\top + Q.$$

This is a discrete-time Riccati recursion of the index t. We can think this recursion as the combination of ascending force from the time update and the decreasing force from the observation update.

Proposition 6. *Given $(F, Q^{1/2})$ stabilizable and (F, H) detectable and $\Sigma_0 \geq 0$, then*

$$\lim_{t\to\infty} \Sigma_{t+1|t} = \Sigma$$

exponentially fast, where Σ is the solution of the discrete-time Riccati equation

$$\Sigma = F\Sigma F^\top + Q - F\Sigma H^\top \left(R + H^\top \Sigma H\right)^{-1} H\Sigma F^\top.$$

Proof. See (5). □

4 Extended Kalman Filter

We now consider a more general dynamic model with time dynamic model:

$$\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{w}_t), \tag{17}$$

and observation model:

$$\mathbf{o}_t = h(\mathbf{s}_t, \mathbf{v}_t). \tag{18}$$

As we have seen in the case with constant inputs, we can approximate these nonlinear models by separating how the mean propagates and how the uncertainty evolves linearly. Let's take time dynamic model as an example.

$$\begin{aligned}
\bar{s}_{t+1} &\approx f(\bar{s}_t, 0), \\
\delta\mathbf{s}_{t+1} &\approx \frac{\partial}{\partial s_t} f(s_t, w_t)\Big|_{s_t=\bar{s}_t, w_t=0} \delta\mathbf{s}_t + \frac{\partial}{\partial w_t} f(s_t, w_t)\Big|_{s_t=\bar{s}_t, w_t=0} \mathbf{w}_t \\
&= F_{s,t}\delta\mathbf{s}_t + F_{w,t}\mathbf{w}_t.
\end{aligned}$$

The matrices $F_{s,t}$ and $F_{w,t}$ are commonly known as the *Jacobian matrices*. They capture how the uncertainty is transferred linearly in this nonlinear model. As $\delta\mathbf{s}_t$ is zero-mean,

$$\Sigma_{t+1|t} \approx F_{s,t}\Sigma_{t|t}F_{s,t}^\top + F_{w,t}QF_{w,t}^\top.$$

Similarly, the observation model can be approximated by

$$\begin{aligned}
\bar{o}_{t+1} &\approx h(\bar{s}_{t+1}, 0), \\
\delta\mathbf{o}_{t+1} &\approx \frac{\partial}{\partial s_{t+1}} h(s_{t+1}, v_{t+1})\Big|_{s_{t+1}=\bar{s}_{t+1}, v_{t+1}=0} \delta\mathbf{s}_t + \frac{\partial}{\partial s_{t+1}} h(s_{t+1}, v_{t+1})\Big|_{s_{t+1}=\bar{s}_{t+1}, v_{t+1}=0} \mathbf{v}_{t+1} \\
&= H_{s,t+1}\delta\mathbf{s}_{t+1} + H_{v,t+1}\mathbf{v}_{t+1}.
\end{aligned}$$

In summary, the time update equations for EKF are given by

$$\begin{aligned}
\bar{s}_{t+1|t} &= f(\bar{s}_t, 0), \\
\Sigma_{t+1|t} &= F_{s,t}\Sigma_{t|t}F_{s,t}^\top + F_{w,t}QF_{w,t}^\top.
\end{aligned}$$

and observation update equations are

$$\begin{aligned}
\bar{s}_{t+1|t+1} &= \bar{s}_{t+1|t} + \Sigma_{t+1|t}H_s^\top \left(H_s\Sigma_{t+1|t}H_s^\top + H_v R H_v^\top\right)^{-1} \left(o_{t+1} - h(\bar{s}_{t+1|t}, 0)\right), \\
\Sigma_{t+1|t+1} &= \Sigma_{t+1|t} - \Sigma_{t+1|t}H_s^\top \left(H_s\Sigma_{t+1|t}H_s^\top + H_v R H_v^\top\right)^{-1} H_s\Sigma_{t+1|t}.
\end{aligned}$$

5 Beyond Kalman Filter

This section provides a roadmap to more advanced topics in Kalman filtering. We can compare to our initial model to see which assumptions are relived in the following topics.

5.1 Kalman-Bucy Filter

The discrete-time system (2) and (3) can be considered as the sample from continuous-time systems, and the filtering method on continuous-time systems is the *Kalman-Bucy* filter. It is a very classical example of stochastic differential equations, but our derivation pretty much covers the essential part in a simpler setting.

5.2 Kalman Filter on a Manifold

Previously, we assume that the state resides in the Euclidean space. However, this assumption rarely holds in many engineering applications. The state of the engineering systems can usually be modeled on a manifold. Therefore, the filtering method on a manifold becomes an important topic, and is still actively studied.

6 Conclusions

I hope this note serves a foundation to understand and to use the Kalman filter, and also provides mathematical insights to interest the readers.

References

[1] K. D. Elworthy, Y. L. Jan, and X.-M. Li, *The Geometry of Filtering*, ser. Frontiers in Mathematics. Birkhäuser Basel, 2010.

[2] Y. Pei, S. Biswas, D. S. Fussell, and K. Pingali, "An elementary introduction to Kalman filtering," *Communications of the ACM*, vol. 62, no. 11, pp. 122–133, Oct. 2019. [Online]. Available: https://doi.org/10.1145/3363294

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, 1st ed. Athena Scientific, 2002.

[4] A. Banerjee, Xin Guo, and Hui Wang, "On the optimality of conditional expectation as a Bregman predictor," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2664–2669, Jul. 2005.

[5] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Pearson, 2000.