# Note 24. Kalman filtering

Alex Fu

Fall 2022

## 1 Introduction

The **Kalman filter** (KF) is an algorithm that tracks an estimate of the state of a stochastic dynamical system, given a sequence of noisy observations of the state over time. It is a recursive, online, and efficient filter: it repeatedly calls itself to work with a stream of observations in real time, with little computational need, and filter out the effects of random noise.

The state estimate of the Kalman filter is linear in the observations, and optimal in minimizing the quadratic cost function of mean squared error, so the Kalman filter is also called the *linear-quadratic estimator* (LQE) and finds itself in fields such as control theory, signal processing, and econometrics. Here, we will present two derivations of the KF, both based on LLSE.

The Kalman filter is very generalizable, but in this note we will work in discrete time, work with time-homogeneous models, and focus on the one-dimensional scalar case.

a. State and observation: $X_n$, $Y_n$, $A$, $C$, $V_n$, $W_n$.

b. Control and feedback*: $U_n$, $B$, $F$.

c. Estimation and error: $\hat{X}_{n|n}$, $\hat{X}_{n|n-1}$, $\sigma^2_{n|n}$, $\sigma^2_{n|n-1}$, $\Sigma_{n|n}$, $\Sigma_{n|n-1}$.

d. Innovation and gain: $\tilde{Y}_n$, $K_n$.

## 2 Algebraic derivation

### 2.1 State and observation

**Definition 1** (State; dynamics; process noise)**.**

The **states** of the dynamical system are random variables $(X_n)_{n\in\mathbb{N}}$, where $X_0$ is usually given, but $(X_n)_{n\geq 1}$ are unknown random variables to be estimated. $X_n$ most commonly describes some sort of position, velocity, or acceleration at time $n$.

The **dynamics** or **transition model** is a scalar $A$ that describes how the state evolves over time, often describing a physical model such as $x = x_0 + vt + \frac{1}{2}at^2$. We assume that $A$ is constant by time-homogeneity, but we may use $(A_n)_{n\in\mathbb{N}}$ if necessary.

The **process noises** are random variables $(V_n)_{n\geq 1}$ assumed to be independent of the states and i.i.d. as $\mathcal{N}(0, \sigma_V^2)$. The precise distribution is less important than the variance $\sigma_V^2$, though when $X_0$ is Gaussian, Kalman filtering with Gaussian noise is exactly the MMSE.

The states, dynamics, and process noises are related by the *state-transition equation*

$$X_n = AX_{n-1} + V_n, \quad n \geq 1.$$

In the vector case, the states $X_n$ and process noises $V_n$ are random vectors in $\mathbb{R}^d$ for $d \geq 1$, and the dynamics is a constant matrix $A \in \mathbb{R}^{d\times d}$.

**Definition 2** (Observation; observation model; observation noise)**.**

The **observations** or **measurements** $(Y_n)_{n\geq 1}$ are random variables known to the algorithm, commonly values taken from sensors or collected data.

The **observation model** is also a scalar $C$ that describes how observations are derived from the true state (deterministically). The **observation noise** or *measurement noise* $(W_n)_{n\geq 1}$, i.i.d. as $\mathcal{N}(0, \sigma_W^2)$ and independent of all other r.v.s, models the uncertainty of measurement.

The observations, observation model, and observation noises are related analogously to the underlying dynamical system by the *state-observation equation*

$$Y_n = CX_n + W_n, \quad n \geq 1.$$

We will assume without loss of generality that $C = 1$, as we can rescale the observations $Y_n$ and the observation noise variance $\sigma_W^2$ if necessary.

In the vector case, the observations $Y_n$ and observation noises $W_n$ are random vectors in $\mathbb{R}^e$, where $e$ does not have to equal $d$. For instance, there could be multiple redundant sensors for a state entry, or no sensors measuring the entry at all. The observation model is a matrix $C \in \mathbb{R}^{e\times d}$.

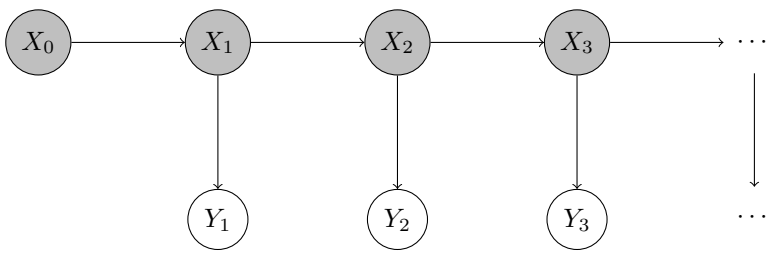


Figure 1: A graphical summary of the states and observations.

Note that the filter assumes the underlying stochastic dynamical system is *linear*. In real systems, nonlinear dynamics might be incorporated into the random process noise instead of the model, which can greatly worsen the performance of the filter.

Other modelling assumptions we leave you to consider: how could we deal with constant *drift* or *offset* in the states? Should we incorporate varying drift into the state or the process noise? How can we describe time-inhomogeneous process or observation noise?

We leave the following exercises in working with one-step recurrence relations and induction.

a. Find $X_n$ in closed form in terms of $A$, $X_0$, and $(V_k)_{k=1}^n$.

b. Find $\mathbb{E}(X_n)$ in terms of $A$ and $\mathbb{E}(X_0)$.

c. Find $\text{var}(X_n)$ in terms of $A$, $\text{var}(X_0)$, and $\Sigma_V$.

d. Find $\lim_{n\to\infty} \text{var}(X_n)$ for a **stable** system, in which $|A| < 1$.

## 2.2 Control and feedback*

**Definition 3** (Control).

A more general state-transition equation includes the random variables of **control inputs** $(U_n)_{n\geq 1}$, and the matrix of the **control-input model** $B$:

$$X_n = AX_{n-1} + BU_n + W_n, \quad n \geq 1.$$

**Definition 4** (Open-loop control; closed-loop control).

There are two broad types of control: the control input $U_n$ is independent of the state $X_{n-1}$ in **open-loop** control, while $U_n$ is some function of $X_{n-1}$ in **closed-loop** or **feedback** control, often a linear function $U_n = FX_{n-1}$. The closed-loop transition model

$$X_n = (A + BF)X_{n-1} + W_n$$

allows for a greater degree of self-correction or self-stabilization of a system.

## 2.3 Estimation and error

**Definition 5** (State estimate).

The goal of the Kalman filter is to track the optimal **estimate** $\hat{X}_n$ of the state $X_n$ at every time step $n \geq 1$, given the *history* or *trajectory* of observations $Y^{(1:n)} = (Y_1, \ldots, Y_n)$:

$$\hat{X}_n = \hat{X}_{n|n} := \underset{f(\cdot)\ \text{affine}}{\operatorname{argmin}}\, \mathbb{E}\Big(\big\|X_n - f(Y^{(1:n)})\big\|^2\Big).$$

In the vector case, the objective function is the more general norm-squared $\mathbb{E}(\|Z\|^2) = \mathbb{E}(Z^\mathsf{T} Z)$. What does the affine function $f(Y^{(1:n)})$ look like for scalar $Y_1, \ldots, Y_n$? What about for random vectors $Y_1, \ldots, Y_n$ with different dimensions from $X_n$?

We can directly find that $\hat{X}_{n|n} = \mathbb{L}(X_n \mid Y^{(1:n)})$, but LLSE requires knowledge about $X_n$ that the algorithm does not have. Instead, we must find $\hat{X}_{n|n}$ using known terms like $A$.

**Definition 6** (State prediction).

The **prediction** of the state $X_n$ at time step $n$, given observations up to time $k < n$, is

$$\hat{X}_{n|k} := \mathbb{L}(X_n \mid Y_1, \ldots, Y_k).$$

We leave the reader to verify that $\hat{X}_{n|k} = A^{n-k}\hat{X}_{k|k}$. The independent, zero-mean process noises "disappear" from the prediction, leaving only the best known estimator $\hat{X}_{k|k}$ "advanced by $n-k$ time steps in the model." Often, the *prediction* will simply refer to $\hat{X}_{n|n-1} = A\hat{X}_{n-1|n-1}$.

**Definition 7** (Estimation variance; prediction variance).

The **estimation variance** at time $n$ is the variance of the estimation residual,

$$\sigma^2_{n|n} = \Sigma_{n|n} := \mathsf{var}(X_n - \hat{X}_{n|n}).$$

We leave the reader to check that $\sigma^2_{n|n} = \mathbb{E}((X_n - \hat{X}_{n|n})^2)$. Defined similarly, the **prediction variance** at time $n$, with observations up to time $k < n$, is

$$\sigma^2_{n|k} = \Sigma_{n|k} := \mathsf{var}(X_n - \hat{X}_{n|k}).$$

**Definition 8** (Estimation error; prediction error).

The **estimation error** at time $n$ is $\mathbb{E}(\|X_n - \hat{X}_{n|n}\|^2)$, the minimum value of the filter's mean squared error cost function. The **prediction error** is defined analogously.

In the scalar case, the estimation variance is precisely the estimation error. In the vector case, the estimation variance is the matrix $\Sigma_{n|n}$, and the estimation error is $\mathbb{E}(\|X_n - \hat{X}_{n|n}{}^2\|) = \mathrm{tr}(\Sigma_{n|n})$. We leave these as exercises in the orthogonality principle and the cyclic property of the trace.

### 2.4 Innovation and gain

Let us now dissect $\hat{X}_{n|n} = \mathbb{L}(X_n \mid Y_1, \ldots, Y_n)$. A first step in finding the LLSE is to orthogonalize $\{1, Y_1, \ldots, Y_n\}$ by the Gram–Schmidt procedure to obtain the **innovations** $\{1, \tilde{Y}_1, \ldots, \tilde{Y}_n\}$:

$$\tilde{Y}_n = Y_n - \mathbb{L}(Y_n \mid Y_1, \ldots, Y_{n-1}) = Y_n - C\hat{X}_{n|n-1}.$$

Note the definition of $Y_n$, $\hat{X}_{n|n-1}$, and the linearity of LLSE estimation.

Now, we can split the projection of $X_n$ onto span $\{1, \tilde{Y}_1, \ldots, \tilde{Y}_n\}$:

$$\begin{aligned}\text{proj}_{\{1,\tilde{Y}_1,\ldots,\tilde{Y}_n\}}(X_n) &= \text{proj}_{\{1,\tilde{Y}_1,\ldots,\tilde{Y}_{n-1}\}}(X_n) + \text{proj}_{\{\tilde{Y}_n\}}(X_n) \\ &= A\hat{X}_{n-1|n-1} + K_n\tilde{Y}_n.\end{aligned}$$

We intentionally split the projection into only two parts so that the filter is *recursive* and *online*. The first term can be found recursively using $\hat{X}_{n-1|n-1}$, and the second term can be found as the new observation $Y_n$ arrives or is made available.

The projection of $X_n$ onto the span of $\tilde{Y}_n$, which is zero-mean, must be some linear transformation of $\tilde{Y}_n$, written $K_n\tilde{Y}_n$. Then the Kalman **gain** at time $n$ is the scalar, or matrix, $K_n$.

We can derive the scalar gain $K_n$ using the formula for an orthogonal projection onto a single zero-mean random variable. Note below that $\hat{X}_{n|n-1}$ is orthogonal to $\tilde{Y}_n$:

$$\begin{aligned}K_n = \frac{\text{cov}(X_n, \tilde{Y}_n)}{\text{var}(\tilde{Y}_n)} &= \frac{\text{cov}(X_n - \hat{X}_{n|n-1}, CX_n + W_n - C\hat{X}_{n|n-1})}{\text{var}(CX_n + W_n - C\hat{X}_{n|n-1})} \\ &= \frac{C\,\text{var}(X_n - \hat{X}_{n|n-1})}{C^2\,\text{var}(X_n - \hat{X}_{n|n-1}) + \text{var}(W_n)} \\ &= \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_W}.\end{aligned}$$

The scalar Kalman gain $0 \leq K_n \leq 1$ can be thus interpreted as a *learning rate*, the proportion of information that can be gained from the new observation $Y_n$. The gain can also be manually tuned to "favor" the existing prediction $\hat{X}_{n|n-1}$ or the innovation $\tilde{Y}_n$.

In the vector case, we find the very similar formula $K_n = \Sigma_{n|n-1}C^\mathsf{T}\left[(C\Sigma_{n|n-1}C^\mathsf{T} + \Sigma_W)^{-1}\right]$.

In summary, the estimate of the true state at time $n$ is a weighted average of the prediction from the previous estimate at time $n-1$ and the new observation at time $n$.

"The optimal estimate of $X_n$ lies between prediction and observation."

$$\hat{X}_{n|n} = \hat{X}_{n|n-1} + K_n\tilde{Y}_n = (I - K_n)\hat{X}_{n|n-1} + K_nY_n$$

Make sure you are comfortable with the derivation above! We leave the following checks:

e. Prove the **orthogonal update**: if $X_n$ is zero-mean, then $\hat{X}_{n|n} = \hat{X}_{n|n-1} + \mathbb{L}(X_n \mid \tilde{Y}_n)$.

f. Find $\mathbb{E}(\hat{X}_{n|n})$ in terms of $A$ and $\mathbb{E}(X_0)$. Hint: you have already found this previously.

g. Find a linear recurrence relation for $\hat{X}_{n|n}$ in terms of only $A$, $K_n$, and $Y_n$.

### 2.5 Prediction and update

The Kalman filter algorithm is most often carried out in two phases: **prediction** and **update**, also called *propagation* and *correction*. The phases typically alternate, but we can also predict several steps in advance without incorporating any new observations, or update several times in sequence to account for multiple newly available observations.

Throughout, we will keep track of the state estimate $\hat{X}_{n|n}$ and the estimation variance $\Sigma_{n|n}$. At initialization, we set $\hat{X}_{0|0} \leftarrow \mathbb{E}(X_0)$ and $\Sigma_{0|0} \leftarrow \mathsf{var}(X_0)$.

In the prediction phase after time step $n-1$, we have access to $(\hat{X}_{n-1|n-1}, \Sigma_{n-1|n-1})$. We can find the *predicted* or *a priori* state estimate and estimation variance:

$$\begin{aligned}
\hat{X}_{n|n-1} &\leftarrow A\hat{X}_{n-1|n-1} \\
\sigma^2_{n|n-1} &= \mathsf{var}(X_n - \hat{X}_{n|n-1}) \\
&= \mathsf{var}(A(X_{n-1} - \hat{X}_{n-1|n-1}) + V_n) \\
&\leftarrow A^2\sigma^2_{n-1|n-1} + \sigma^2_V
\end{aligned}$$

Interestingly, we can already find the Kalman gain here:

$$K_n \leftarrow \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_W}.$$

In the update phase at time step $n$, we have found $(\hat{X}_{n|n-1}, \Sigma_{n|n-1})$, and the new observation $Y_n = CX_n + W_n$ becomes available. We can then find the innovation, and the *a posteriori* state estimate and estimate variance:

$$\begin{aligned}
\tilde{Y}_n &\leftarrow Y_n - C\hat{X}_{n|n-1} \\
\hat{X}_{n|n} &\leftarrow \hat{X}_{n|n-1} + K_n\tilde{Y}_n \\
\sigma^2_{n|n} &= \mathsf{var}(X_n - [(I - K_nC)\hat{X}_{n|n-1} + K_nY_n]) \\
&= \mathsf{var}((I - K_nC)\cdot(X_n - \hat{X}_{n|n-1}) - K_nW_n) = \cdots \\
&\leftarrow (I - K_n)\cdot\sigma^2_{n|n-1}.
\end{aligned}$$

The algorithm is quite space-efficient: it does not need to store any past estimates or observations. It is also time-efficient: only the computation of $\hat{X}_{n|n}$ is online. The Kalman gains $K_n$, estimation variances $\sigma^2_{n|n}$, and prediction variances $\sigma^2_{n|n-1}$ can all be recursively found and stored *offline*.

## 3 Geometric derivation

We can also derive the scalar Kalman filter equations by leveraging geometry in the Hilbert space of random variables. Make sure you know which orthogonality relations hold — the dimension of span $\{1\}$ is omitted, and not all orthogonal projections are drawn vertically! These diagrams are visualizations of infinite-dimensional subspaces after all.

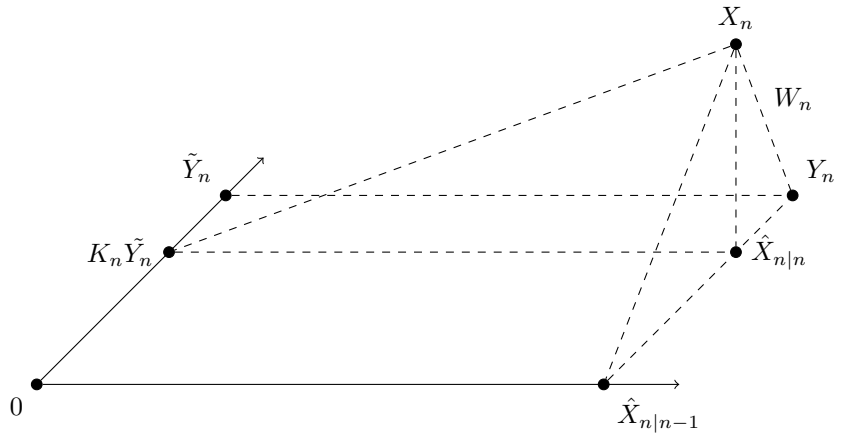


Figure 2: $\hat{X}_{n|n}$ is the orthogonal projection of $X_n$ onto span $\{1, Y_1, \ldots, Y_{n-1}\} \oplus$ span $\{\tilde{Y}_n\}$.

Let us first see what the length of the only term involving $K_n$ is equal to:

$$\left\|K_n\tilde{Y}_n\right\| = \left\|\hat{X}_{n|n} - \hat{X}_{n|n-1}\right\| = K_n\left\|Y_n - \hat{X}_{n|n-1}\right\|.$$

Now, we can leverage the similiarity between two triangles: the "smaller" $(\hat{X}_{n|n-1}, \hat{X}_{n|n}, X_n)$ and the "larger" $(\hat{X}_{n|n-1}, X_n, Y_n)$, both of whose hypotenuses are known.

$$\begin{aligned} K_n = \frac{\left\|\hat{X}_{n|n} - \hat{X}_{n|n-1}\right\|}{\left\|Y_n - \hat{X}_{n|n-1}\right\|} &= \frac{\left\|\hat{X}_{n|n} - \hat{X}_{n|n-1}\right\|}{\left\|X_n - \hat{X}_{n|n-1}\right\|}\frac{\left\|X_n - \hat{X}_{n|n-1}\right\|}{\left\|Y_n - \hat{X}_{n|n-1}\right\|} \\ &= \left(\frac{\left\|X_n - \hat{X}_{n|n-1}\right\|}{\left\|Y_n - \hat{X}_{n|n-1}\right\|}\right)^2 \\ &= \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_W}. \end{aligned}$$

By the Pythagorean theorem applied to the "smaller" triangle, we can also find

$$\begin{aligned} \sigma^2_{n|n} = \left\|X_n - \hat{X}_{n|n}\right\|^2 &= \left\|X_n - \hat{X}_{n|n-1}\right\|^2\left(1 - \frac{\left\|\hat{X}_{n|n} - \hat{X}_{n|n-1}\right\|^2}{\left\|X_n - \hat{X}_{n|n-1}\right\|^2}\right) \\ &= (1 - K_n)\cdot\sigma^2_{n|n-1}. \end{aligned}$$

In order to find $\sigma^2_{n|n-1}$, we will need to draw $X_{n-1}$, which introduces a possible new dimension. The following diagram is slightly "rotated towards the reader" from the diagram above; note the different "vertical" projections in the two diagrams.

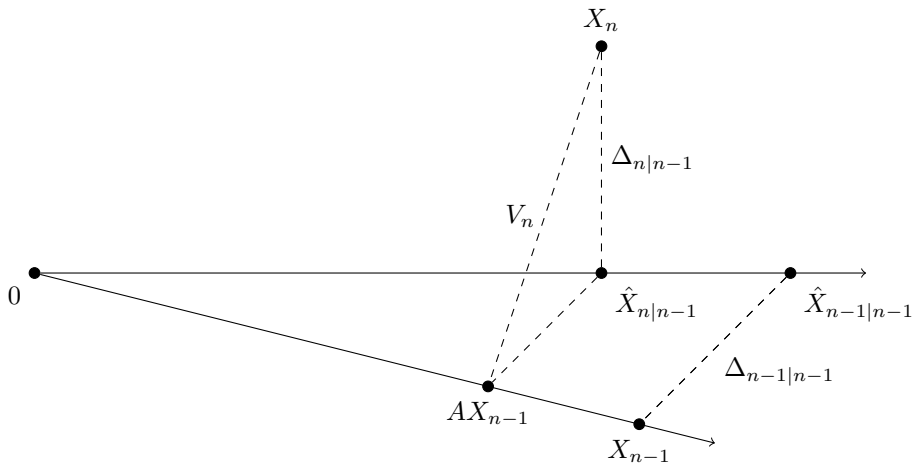


Figure 3: The difference between $AX_{n-1}$ and the prediction $\hat{X}_{n|n-1}$ is orthogonal to the noise $V_n$.

Applying the Pythagorean theorem to the triangle formed by $(AX_{n-1}, \hat{X}_{n|n-1}, X_n)$, and leveraging the similarity of the other two triangles, we find that

$$\begin{aligned}\sigma^2_{n|n-1} = \left\|\Delta_{n|n-1}\right\|^2 &= \left\|AX_{n-1} - A\hat{X}_{n-1|n-1}\right\|^2 + \|V_n\|^2 \\ &= A^2\sigma^2_{n-1|n-1} + \sigma^2_V.\end{aligned}$$

## 4 Summary

For time steps $n \geq 1$, the states and observations are given by the following equations.

$$\begin{aligned}X_n &= AX_{n-1} + V_n \\ Y_n &= CX_n + W_n.\end{aligned}$$

We initialize the state estimate and estimation variance as $(\hat{X}_{0|0}, \Sigma_{0|0}) \leftarrow (\mathbb{E}(X_0), \mathsf{var}(X_0))$. The Kalman gains and estimation variances can be found offline as follows.

$$\begin{aligned}\sigma^2_{n|n-1} &= A^2\sigma^2_{n-1|n-1} + \sigma^2_V && \text{(prediction)} \\ K_n &= \sigma^2_{n|n-1}(\sigma^2_{n|n-1} + \sigma^2_W)^{-1} && \text{(gain)} \\ \sigma^2_{n|n} &= (I - K_n) \cdot \sigma^2_{n|n-1} && \text{(update)}\end{aligned}$$

The state estimates are updated online as new observations arrive.

$$\begin{aligned}\hat{X}_{n|n-1} &= A\hat{X}_{n-1|n-1} && \text{(prediction)} \\ \tilde{Y}_n &= Y_n - \hat{X}_{n|n-1} && \text{(innovation)} \\ \hat{X}_{n|n} &= \hat{X}_{n|n-1} + K_n\tilde{Y}_n && \text{(update)}\end{aligned}$$

■