Kalman Filtering

1. Motivation:

So far, we have considered how best to record data from measurement systems and how to analyze the spectral properties of those measurements. One such measurement system from which we might record might be represented by Figure 1.

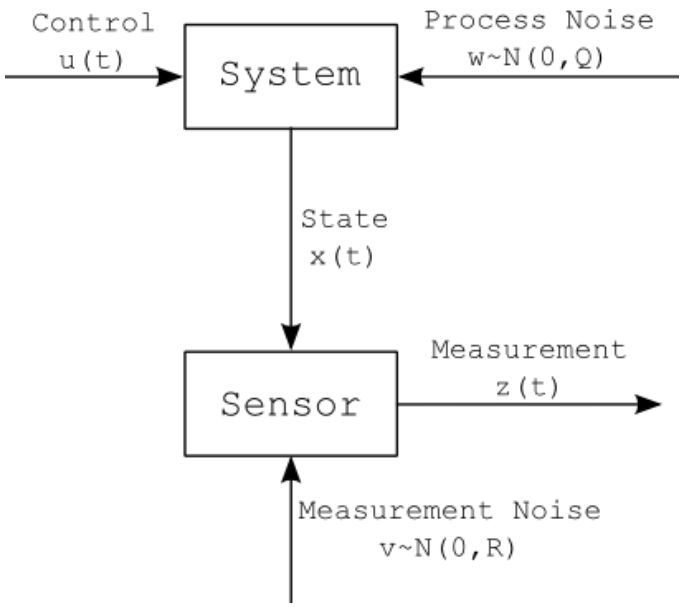

Figure 1. Schematic of a measurement system, with optional input for control applications. Key to note is that the only output is the measurement, z(t), with the true state of the system, x(t), hidden from view of the experimenter.

Normally, one would take the output, z(t), of the system described in Figure 1 and then go along happily analyzing it as though it suitably represented their system. But let's get metaphysical for a moment; our measurement, z(t), is ultimately some noisy transformation of the inaccessible "true" system state, x(t). What do we do if z(t) is very noisy?

In light of the messiness that data can exhibit, one might turn to a model-based approach. However, a model is a very clean, tidy, and ultimately over-simplified representation of the process, and as such removes the influences of latent variables that give rise to our process noise. While a simplified description of a system is often sufficient and occasionally exemplary, in a situation where we care less about succinctness and more about accurately representing x(t), the model-based approach will fail to capture the fluctuations present in an actual system.

If models alone do not work, and measurements are too noisy, is estimating x(t) simply a lost cause? Not necessarily! We can combine model and measurement, prior and posterior, to come up with an optimal estimate of x(t). The Kalman filter is precisely this estimator.

2. The Bayesian estimator:

The Kalman filter is a class of Bayesian estimator. You may be familiar with Bayes' theorem:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \quad (1)$$

In the framework of hypothesis testing and estimation, $p(\theta)$ is the prior distribution (the likelihood of each hypothesis being tested, independent of any data), $p(x|\theta)$ is the likelihood of some given observations conditioned upon each hypothesis, $p(\theta|x)$ is known as the posterior, i.e., the likelihood of each hypothesis conditioned upon the data, and $p(x)$ is essentially a scaling factor.

While the Kalman filter, in practice, is not a straightforward application of Bayes' theorem, it is derived using Bayes' theorem and, as with all Bayesian estimators, is proven to uniformly be the optimal minimum mean squares estimator (MMSE) for x(t) in Figure 1, at least in a linear sense (LMMSE). Essentially, a MMSE is an estimator that aims to minimize the expected square error, $E[(x-\hat{x})^2]$, between the estimate, $\hat{x}$, and the quantity being estimated, x. A LMMSE is a MMSE obtained exclusively through linear operations.

3. Formulating the problem

Enough abstraction, let's make our system explicit. We will model our state value x(n) (notice we are now talking about a discrete variable) as arising from the following matrix update rule:

$$\mathbf{x_n} = \mathbf{A}\mathbf{x_{n-1}} + \mathbf{B}\mathbf{u_n} + \{\mathbf{w}_n \sim \mathcal{N}(0, \mathbf{Q}_n)\} \quad (2)$$

where $\mathbf{A}$ is the state transition matrix, $\mathbf{x_n}$ is the state vector at step n, $\mathbf{B}$ and $\mathbf{u_n}$ are optional control parameters (which will be omitted for the rest of this handout), and $\mathbf{w_n}$ is a random vector drawn from a normal distribution parameterized by zero mean and covariance matrix $\mathbf{Q_n}$. We will then model our measurement of this process as follows:

$$\mathbf{z_n} = \mathbf{H}\mathbf{x_n} + \{\mathbf{v}_n \sim \mathcal{N}(0, \mathbf{R}_n)\} \quad (3)$$

where $\mathbf{z_n}$ is the measurement at step n, $\mathbf{H}$ is the transduction matrix, and $\mathbf{v_n}$ is a random vector drawn from a normal distribution parameterized by zero mean and covariance matrix $\mathbf{R_n}$.

4. Assessing assumptions and essential properties

Before we continue, we must observe our system and identify what assumptions and constraints we are imposing on our variables. In particular, equations 2 and 3 describe the Gauss-(Hidden) Markov model for the time evolution of an observed random variable and an underlying unobserved random variable. With this model for state evolution, we assume that $\mathbf{w_n}$ and $\mathbf{v_n}$ are GWN processes such that

$$\forall m: \forall n \neq m: \{E[\mathbf{w_m}\mathbf{w_n}^T] = 0 \cap E[\mathbf{v_m}\mathbf{v_n}^T] = 0\} \tag{4}$$

In other words, for each time point n, the value drawn for $\mathbf{w_n}$ and $\mathbf{v_n}$ is independent of any other values drawn for these processes at any other time. Furthermore, we assume that $\mathbf{w_n}$ and $\mathbf{v_n}$ are both independent of the initial state vector, $\mathbf{x_0}$. Given these assumptions, we know the process is Markov (i.e., dependent only on the directly preceding time lag) There is a rigorous proof for this property, but it's laborious and a bit of a distraction, so you'll have to trust me. This is an important property, as it then allows us to iteratively apply the model to generate estimates for all n.

5. Deriving the estimate

With the tedium of verifying nice statistical properties out of the way, let us now define the estimates that we will be taking. We will be making estimates iteratively for each time n, but to start, let us consider some specific n after we have already obtained its prior estimate, which we shall designate as $\hat{\mathbf{x}}_\mathbf{n}^-$. From here, we define the following error vectors for our existing prior estimate and our soon-to-exist posterior estimate:

$$\mathbf{e}_\mathbf{n}^- \coloneqq \mathbf{x_n} - \hat{\mathbf{x}}_\mathbf{n}^- \tag{5}$$

$$\mathbf{e}_\mathbf{n}^+ \coloneqq \mathbf{x_n} - \hat{\mathbf{x}}_\mathbf{n}^+ \tag{6}$$

In other words, we define how far off our estimates are w.r.t. the unseen variable $\mathbf{x_n}$. Do not worry that we cannot evaluate this; we will be able to evaluate after some transformations and definitions to follow. To continue, we may now consider the error covariance matrices to be defined as such:

$$\mathbf{P}_\mathbf{n}^- = E(\mathbf{e}_\mathbf{n}^-\mathbf{e}_\mathbf{n}^{-T}) \tag{7}$$

$$\mathbf{P}_\mathbf{n}^+ = E\left(\mathbf{e}_\mathbf{n}^+\mathbf{e}_\mathbf{n}^{+T}\right) \tag{8}$$

We now seek to identify a way to incorporate our most recent measurement, $\mathbf{z_n}$, into our prior estimate in order to obtain a posterior estimate. This is where the linear "L" in our LMMSE comes into play; we will simply take a term called the "innovation", which corresponds to the difference between our measurement and post-transduction prior estimate, then apply a gain to it and add it to our prior to obtain our posterior. In math terms:

$$\hat{\mathbf{x}}_\mathbf{n}^+ = \hat{\mathbf{x}}_\mathbf{n}^- + \mathbf{K}(\mathbf{z_n} - \mathbf{H}\hat{\mathbf{x}}_\mathbf{n}^-) \tag{9}$$

This equation may seem to be a bit ad hoc, but it actually arises from applying our measurement formula in Equation 3 to a Bayesian formulation (as in Equation 1) where θ corresponds to x and x (confusingly) corresponds to the set of all measurements from step 1 to n. The derivation is, as with the Markov property, quite tedious, so I will not be showing it today.

From here, we need to find the optimal gain $\mathbf{K}$. We can start by making substitutions. By subbing (6) in for (8), we obtain

$$\mathbf{P}_\mathbf{n}^+ = \mathrm{E}\left((\mathbf{x_n} - \hat{\mathbf{x}}_\mathbf{n}^+)(\mathbf{x_n} - \hat{\mathbf{x}}_\mathbf{n}^+)^\mathrm{T}\right) \tag{10}$$

We can also substitute (3) into (9) to obtain

$$\hat{\mathbf{x}}_\mathbf{n}^+ = \hat{\mathbf{x}}_\mathbf{n}^- + \mathbf{K}(\mathbf{H}\mathbf{x_n} + \mathbf{v}_\mathrm{n} - \mathbf{H}\hat{\mathbf{x}}_\mathbf{n}^-) \tag{11}$$

which in turn can be substituted into (10) to obtain

$$\mathbf{P}_\mathbf{n}^+ = \mathrm{E}(\{\mathbf{x_n} - [\hat{\mathbf{x}}_\mathbf{n}^- + \mathbf{K}(\mathbf{H}\mathbf{x_n} + \mathbf{v}_\mathrm{n} - \mathbf{H}\hat{\mathbf{x}}_\mathbf{n}^-)]\}\{\mathbf{x_n} - [\hat{\mathbf{x}}_\mathbf{n}^- + \mathbf{K}(\mathbf{H}\mathbf{x_n} + \mathbf{v}_\mathrm{n} - \mathbf{H}\hat{\mathbf{x}}_\mathbf{n}^-)]\}^\mathrm{T}) \tag{12}$$

From here, I will go on assuming that all matrices and vectors are now scalars to avoid any tricky linear algebra while still giving a flavor of the derivation, but I will show the matrix representation of the Kalman gain at the end to show similarities between the univariate and multivariate cases. To continue with the scalar formalization, we now abuse the linearity of the expectation operation to collect and distribute expectation terms:

$$\mathbf{P}_\mathbf{n}^+ = \mathrm{E}[(\mathbf{x_n} - \hat{\mathbf{x}}_\mathbf{n}^-)^2] - 2\mathbf{K}\mathrm{E}\{[\mathbf{H}(\mathbf{x_n} - \hat{\mathbf{x}}_\mathbf{n}^-) + \mathbf{v}_\mathrm{n}](\mathbf{x_n} - \hat{\mathbf{x}}_\mathbf{n}^-)\} + \mathbf{K^2}\mathrm{E}\{[\mathbf{H}(\mathbf{x_n} - \hat{\mathbf{x}}_\mathbf{n}^-) + \mathbf{v}_\mathrm{n}]^\mathbf{2}\} \tag{13}$$

$$\mathbf{P}_\mathbf{n}^+ = \mathbf{P}_\mathbf{n}^- - 2\mathbf{K}\mathbf{H}\mathbf{P}_\mathbf{n}^- - 2\mathbf{K}\mathrm{E}\{\mathbf{v}_\mathrm{n}(\mathbf{x_n} - \hat{\mathbf{x}}_\mathbf{n}^-)\} + \mathbf{K^2}\mathbf{H^2}\mathbf{P}_\mathbf{n}^- + \mathbf{2K^2}\mathbf{H}\mathrm{E}\{(\mathbf{x_n} - \hat{\mathbf{x}}_\mathbf{n}^-)\mathbf{v}_\mathrm{n}\} + \mathbf{K^2}\mathbf{R_n} \tag{14}$$

From here, we abuse independence of $\mathbf{v_n}$ and $\mathbf{e_n^-}$ to obtain

$$\mathbf{P_n^+ = P_n^- - 2KHP_n^- + K^2H^2P_n^- + K^2R_n} \tag{15}$$

It should be noted that $\mathbf{P_n^+}$, the variance of the state centered about our estimate, is equivalent to the MMSE term that we wish to optimize. Therefore, to find the optimal gain, we differentiate with respect to $\mathbf{K}$, set the derivative to zero to find the extremum, and solve for $\mathbf{K}$. This gives

$$\begin{gathered}\frac{\mathrm{d}\mathbf{P_n^+}}{\mathrm{d}\mathbf{K}} = -2\mathbf{HP_n^-} + 2\mathbf{KH^2P_n^-} + 2\mathbf{KR_n} \\ \mathbf{HP_n^-} = \mathbf{KH^2P_n^-} + \mathbf{KR_n} \\ \frac{\mathbf{HP_n^-}}{\mathbf{H^2P_n^- + R_n}} = \mathbf{K}\end{gathered} \tag{16}$$

From here, we now go back to (15) and solve for $\mathbf{P_n^+}$:

$$\begin{gathered}\mathbf{P_n^+ = P_n^- - 2KHP_n^- + K^2(H^2P_n^- + R_n)} \\ \mathbf{P_n^+ = P_n^- - 2KHP_n^- + K^2}\frac{\mathbf{HP_n^-}}{\boldsymbol{K}} \\ \mathbf{P_n^+ = P_n^-(1 - KH)}\end{gathered} \tag{17}$$

Compare these results to the following matrix formulations of the equations and note the similarities:

$$\mathbf{P_n^-H^T(HP_n^-H^T + R_n)^{-1} = K} \tag{18}$$

$$\mathbf{P_n^+ = (I - KH)P_n^-} \tag{19}$$

Also note that the gain $\mathbf{K}$ can and will change with each step; the subscript was merely omitted for some semblance of notational simplicity.

At this point, we have finished the estimation phase and are now looking to establish a prior for the next step. This is achieved by passing our estimate, $\hat{\mathbf{x}}_\mathbf{n}^+$ (from equation 9), the closest thing we have to the "true" state vector $\mathbf{x_n}$, into our system equation (2):

$$\hat{\mathbf{x}}_{\mathbf{n+1}}^- = \mathbf{A}\{\hat{\mathbf{x}}_{\mathbf{n+1}}^- \sim \mathcal{N}(\hat{\mathbf{x}}_\mathbf{n}^+, \mathbf{P_n^+})\} + \{\mathbf{w_n} \sim \mathcal{N}(0, \mathbf{Q}_{n+1})\} \tag{20}$$

$$\hat{\mathbf{x}}_{\mathbf{n+1}}^- \coloneqq \mathrm{E}[\hat{\mathbf{x}}_{\mathbf{n+1}}^-] = \mathbf{A}\hat{\mathbf{x}}_\mathbf{n}^+ \tag{21}$$

$$\mathbf{P_{n+1}^-} \coloneqq \mathrm{E}\left[\hat{\mathbf{x}}_{\mathbf{n+1}}^- \hat{\mathbf{x}}_{\mathbf{n+1}}^{-\mathrm{T}}\right] = \mathbf{AP_n^+A^T} + \mathbf{Q}_{n+1} \tag{22}$$

From here, we rinse and repeat.

6. Implementing the estimator

Now that we have our equations for the Kalman filter, we can go ahead and implement it! Before we start turning the Bayesian crank, however, we need to establish our set of priors. Table 1 gives the information one first needs in order to implement the Kalman filter.

Table 1. Required variables for implementing Kalman filter

Variable	Explanation
$\mathbf{A}$	State transition matrix for model
$\mathbf{H}$	Transduction matrix for measurement
$\{\mathbf{Q}_n\}$	Process noise covariance matrices
$\{\mathbf{R}_n\}$	Measurement noise covariance matrices
$\{\mathbf{z}_n\}$	Sequence of measurements over time
$\hat{\mathbf{x}}_0^-$	Initial prior estimate of state vector
$\mathbf{P}_0^-$	Initial prior estimate of state covariance

The first two items, $\mathbf{A}$ and $\mathbf{H}$, describe the model and therefore are crucial to know beforehand. An implicit assumption of the Kalman filter is that the model is a decent one; poor models will actually perform worse than raw data in terms of tracking the "true" state $\mathbf{x}_n$.

The sequences of covariance matrices, $\{\mathbf{Q}_n\}$ and $\{\mathbf{R}_n\}$, can also be difficult to estimate, especially if they are nonstationary and therefore need to be uniquely defined for each step n. $\{\mathbf{Q}_n\}$ is often particularly difficult to estimate, whereas $\{\mathbf{R}_n\}$ can often be evaluated assuming common noise sources (e.g., thermal noise in electrodes) or by simply taking a dummy recording of the environment. Oftentimes, these parameters must be explicitly tuned on some test data before they can be released into the wild.

The sequence of measurements over time, $\{\mathbf{z}_n\}$, need not be available all at once, and in fact it goes against the iterative spirit of the Kalman filter to do so. The Kalman filter is often used for efficient real-time tracking; efficient in the sense that it only requires the most recent data point be stored in memory.

The initial prior estimates are things that ultimately need to be set somewhat arbitrarily by the user.

Below, in Figure 2, is a figure depicting the iterative algorithm and where to apply the necessary equations.

Estimation Step:

$$\mathbf{K} = \mathbf{P_n^-}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P_n^-}\mathbf{H}^{\mathrm{T}} + \mathbf{R_n})^{-1}$$
$$\hat{\mathbf{x}}_\mathbf{n}^+ = \hat{\mathbf{x}}_\mathbf{n}^- + \mathbf{K}(\mathbf{z_n} - \mathbf{H}\hat{\mathbf{x}}_\mathbf{n}^-)$$
$$\mathbf{P_n^+} = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P_n^-}$$

Prediction Step:

$$\hat{\mathbf{x}}_{\mathbf{n+1}}^- = \mathbf{A}\hat{\mathbf{x}}_\mathbf{n}^+ \ (+\boldsymbol{B}\boldsymbol{u}_{n+1})$$
$$\mathbf{P_{n+1}^-} = \mathbf{A}\mathbf{P_n^+}\mathbf{A}^{\mathrm{T}} + \mathbf{Q}_{n+1}$$

LOOP BACK!

7. Demonstration

- Go over the script Kalman_Vm
- Change the sw and sv parameters to show off how it behaves at the limits

8. Advanced Topics

Parameter estimation

- Augmented state vector
- Solve for parameter as yet another state vector element
- Likely takes form of optimal control parameter u(t)

Nonlinear estimation

- Linearize equation by taking Taylor expansion, truncating at first term
- This is known as the Extended Kalman Filter
- No longer guaranteed to be MMSE, merely an ad hoc workaround

Appendix A. Sketch of proof of Markov property

To prove this, we determine the probability densities $p(\mathbf{x_n})$ and $p(\mathbf{x_{n-1}})$, then also evaluate $p(\mathbf{x_n}|\mathbf{x_{n-1}})$, then use the Chapman-Kolmogorov equation to determine if $p(\mathbf{x_n})$ is equal to $\int p(\mathbf{x_n}|\mathbf{x_{n-1}})p(\mathbf{x_{n-1}})d\mathbf{x_{n-1}}$ See Nahum Shimkin's 2009 lecture notes pp.2-3 for more info.

Appendix B. Sketch of justification for Equation 9

Again, integrals need to be evaluated, it's laborious. There is one important transformation, however, whereby we set our prior estimate, $p(\mathbf{x_n})$, to be $\mathbf{x_n}$ conditioned on all measurements up to (but not including) $\mathbf{z_n}$, and also set $p(\mathbf{z_n})$ and $p(\mathbf{z_n}|\mathbf{x_n})$ as conditioned upon all previous measurements. This then allows one to use the distribution of the prior estimate and the measurement model to evaluate the relevant distributions for any given time step, which can then be iterated due to the Markov property. See Brown and Hwang 1997 pp.228-231 for more info.