Gaussian RVs and the Kalman filter

ROB 501

Necmiye Ozay

- **Gaussian random variables**
  - **MVE another look (interpretation based on conditional probability using Schur complements)**
- **(if time) Derivation of Kalman filter and why it is a minimum variance estimator**

Heart of the recursion in Kalman filter

Definition of Terms:

$$\widehat{x}_{k|k} := \mathcal{E}\{x_k|y_0,\cdots,y_k\}$$ ← conditional mean at time k

$$P_{k|k} := \mathcal{E}\{(x_k-\widehat{x}_{k|k})(x_k-\widehat{x}_{k|k})^T|y_0,\cdots,y_k\}$$ ← conditional cov at time k

$$\widehat{x}_{k+1|k} := \mathcal{E}\{x_{k+1}|y_0,\cdots,y_k\} = \mathcal{E}\{A_k x_k + G_k w_k | y_0,\cdots y_k\}$$

$$P_{k+1|k} := \mathcal{E}\{(x_{k+1}-\widehat{x}_{k+1|k})(x_{k+1}-\widehat{x}_{k+1|k})^T|y_0,\cdots,y_k\}$$

$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$, $X \sim x_k$, $Y \sim y_k$, $Z \sim \{y_0,\cdots,y_{k-1}\}$

We want to compute $X \mid \begin{bmatrix} Y \\ Z \end{bmatrix}$ (in Kalman filter $\hat{x}_{k|k}$)

→ for recursive computation, we want to express $X \mid \begin{bmatrix} Y \\ Z \end{bmatrix}$ using $X \mid Z$ (which is $\hat{x}_{k|k-1}$) and Y

$$x_{k|k-1} = A_{k-1}x_{k-1} + G_{k-1}w_{k-1}$$

$\hat{x}_{k|k} = E\{X \mid \begin{bmatrix} Y \\ Z \end{bmatrix}\}$

$$X \mid \begin{bmatrix} Y \\ Z \end{bmatrix} \sim \frac{f_{XYZ}}{f_{YZ}}$$ ↳ joint distribution of X, Y, Z

$$\frac{f_{XYZ}}{f_Z} \sim \begin{bmatrix} X \\ Y \end{bmatrix} \mid Z = \begin{bmatrix} X|Z \\ Y|Z \end{bmatrix} \sim f_{(X|Z),(Y|Z)}$$ (*)

↳ joint distribution of (X|Z) and (Y|Z)

$$(X|Z) \mid (Y|Z) \sim \frac{f_{(X|Z),(Y|Z)}}{f_{(Y|Z)}}$$ (*)

$$= \frac{f_{XYZ}/f_Z}{f_{YZ}/f_Z}$$

$$= \frac{f_{XYZ}}{f_{YZ}} \sim X \mid \begin{bmatrix} Y \\ Z \end{bmatrix}$$

(something we can compute recursively)

batch version

Rob 501 Handout: Grizzle
Useful Facts About Gaussian Random Variables and Vectors

Def. A random variable X is normally distributed with mean μ and variance $\sigma^2 > 0$ if it has density

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The standard deviation is $\sigma > 0$. The mean and variance satisfy

$$\mu := \mathcal{E}\{X\} := \int_{\mathbb{R}} x f_X(x) dx := \int_{-\infty}^{\infty} x f_X(x) dx$$

$$\sigma^2 := \mathcal{E}\{(X-\mu)^2\} := \int_{\mathbb{R}} (x-\mu)^2 f_X(x) dx := \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx$$

Remarks: You should be quite familiar with the "bell curve". X is also called a Gaussian random variable. We often say X has a *univariate normal distribution* or a *univariate Gaussian distribution* to emphasize that we are talking about a single random variable.

For the most part, we do not care too much about individual random variables. We are interested in collections of random variables and random vectors, and hence we are primarily concerned about *jointly distributed random variables*. If you take EECS 501, you can learn a tremendous amount of material about this subject. In the following, I will give a bare bones accounting of *multivariate normal random variables*.

Def. A finite collection random variables $X_1, X_2, \cdots, X_p$, or equivalently, the random vector

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

has a (non-degenerate) *multivariate normal distribution* with mean μ and covariance $\Sigma > 0$ if the joint density is given by

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}.$$

Remark: In the above, $|\Sigma| = \det(\Sigma)$, which must be non-zero for the denominator to be well defined. This condition is what is meant by "non-degenerate". When $|\Sigma| = 0$, one can still define a multivariate normal distribution, but the "moment generating function" must be used. This is a technicality that we will skip.

$$\mathcal{E}\{X\} = \mu \in \mathbb{R}^p \quad \mu_i := \int_{\mathbb{R}^p} x_i \, f_X(x) dx := \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i \, f_X(x_1, \cdots, x_p) dx_1 \cdots dx_p$$

$$\mathrm{cov}(X, X) = \mathcal{E}\{(X - \mu)(X - \mu)^\top\} = \Sigma \in \mathbb{R}^{p \times p}$$

$$\mathcal{E}\{(X_i - \mu_i)(X_j - \mu_j)\} = [\Sigma]_{ij} =: \Sigma_{ij} := \int_{\mathbb{R}^p} (x_i - \mu_i)(x_j - \mu_j) \, f_X(x) dx$$

$$x = (x_1, x_2, \cdots, x_p) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \quad \text{(depending on context)}$$

Marginal Distributions: Each random variable X_i has a *univariate normal distribution* with mean μ_i and variance Σ_{ii}.

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}}.$$

Remark: We note the unfortunate lack of coordination of notation in that the standard deviation of X_i, which we typically denote by σ_i, is given by

$$\sigma_i = \sqrt{\Sigma_{ii}}.$$

I guess we will not be denoting the entries of Σ with lower case σ.

Independence: Gaussian random variables are very special in that they are independent if, and only if, they are uncorrelated. Hence, X_i and X_j are *independent* if, and only if, $\Sigma_{ij} = \Sigma_{ji} = 0$.

Linear Combinations: Define a new random vector by $Y = AX + b$. Then Y is a Gaussian (normal) random vector with

$$\mathcal{E}\{Y\} = A\mu + b =: \mu_Y$$

$$\operatorname{cov}(Y,Y) = \mathcal{E}\{(Y-\mu_Y)(Y-\mu_Y)^\top\} = A\Sigma A^\top =: \Sigma_{YY}$$

Indeed, $Y - \mu_Y = A(X-\mu)$. Hence,

$$\operatorname{cov}(Y,Y) = \mathcal{E}\{[A(X-\mu)][A(X-\mu)]^\top\} = A\mathcal{E}\{(X-\mu)(X-\mu)^\top\}A^\top = A\Sigma A^\top.$$

Remark: Taking $b = 0$ and A to be a row vector with all zeros except a one in the i-th spot, that is $A = [0, \cdots, 1, \cdots, 0]$, recovers the *marginal* distributions discussed above.

Working with Two Vectors of Gaussian Random Variables: In addition to looking at individual random variables making up a random vector, we can group the components to form two or more blocks of vectors as long as their sizes add up to p, the number of components in X. We abuse notation and write

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{matrix} \in \mathbb{R}^n \\ \in \mathbb{R}^m \end{matrix}$$

In books, you'll often see the blocks expressed in bold font, such as $\mathbf{X_1}$ and $\mathbf{X_2}$. We will NOT do this. Conformally with this decomposition of X into two blocks, we decompose the mean and covariance as follows

$$\mu =: \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

and

$$\Sigma =: \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Remark: From our results on the Schur complement, we know that $\Sigma > 0$ if, and only if, $\Sigma_{22} > 0$ and $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} > 0$.

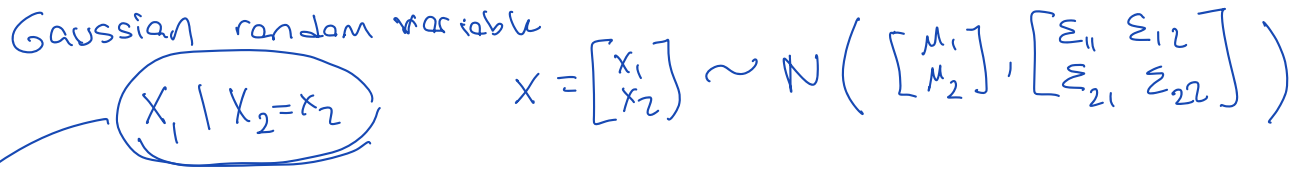
To be super clear on the dimensions, we suppose $n + m = p$ and note that

$$\mu_1 = \mathcal{E}\{X_1\} \in \mathbb{R}^n \quad \text{and} \quad \mu_2 = \mathcal{E}\{X_2\} \in \mathbb{R}^m$$
$$\text{cov}(X_1, X_1) = \Sigma_{11} \in \mathbb{R}^{n \times n} \quad \text{and} \quad \text{cov}(X_2, X_2) = \Sigma_{22} \in \mathbb{R}^{m \times m}$$
$$\text{cov}(X_1, X_2) = \Sigma_{12} \in \mathbb{R}^{n \times m} \quad \text{and} \quad \text{cov}(X_2, X_1) = \Sigma_{21} \in \mathbb{R}^{m \times n}$$

Furthermore, because $\Sigma = \Sigma^\top$, we have that

$$\Sigma_{11}^\top = \Sigma_{11}, \quad \Sigma_{22}^\top = \Sigma_{22}, \text{ and } \Sigma_{12}^\top = \Sigma_{21}.$$

Remark: Each vector X_i has a multivariate normal distribution with mean μ_i and covariance Σ_{ii}. This is also called the **marginal distribution** of X_i. If we know the mean and covariance for the composite vector X, it is very easy to read off the marginal distributions of its vector components.



Fact 1: Conditional Distributions of Gaussian Random Vectors: Let X_1 and X_2 be as above, namely they are components of a larger vector X that has a multivariate normal distribution. Then the conditional distribution of X_1 given $X_2 = x_2$ has a multivariate normal distribution with

$$\text{Mean}: \quad \mu_{1|2} := \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

and

$$\text{Covariance}: \quad \Sigma_{1|2} := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

In passing, we note that the mean depends on the value of x_2 while the covariance does not.

To be extra clear on the meanings here,

- $\mu_{1|2} = \mathcal{E}\{X_1 \mid X_2 = x_2\}$
- $\Sigma_{1|2} = \mathcal{E}\{(X_1 - \mu_{1|2})(X_1 - \mu_{1|2})^\top \mid X_2 = x_2\}$
- X_1 given $X_2 = x_2$ is a random vector. It has a multivariate normal distribution with the above mean vector and covariance matrix. Specifically, its density is

$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} = \frac{1}{\sqrt{(2\pi)^n |\Sigma_{1|2}|}} e^{-\frac{1}{2}(x_1-\mu_{1|2})^\top \Sigma_{1|2}^{-1}(x_1-\mu_{1|2})},$$

  where it is emphasized that $\mu_{1|2}$ depends explicitly on x_2.

Remark: A proof of this can be found at the link below. The algebra is rather painful. If you are very ambitious, you can work out the special case where X_1 and X_2 are scalars. This will not be on any exam.
`http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html`
See also
`http://www.stats.ox.ac.uk/~steffen/teaching/bs2HT9/gauss.pdf`.

Fact 2 on Conditional Independence: Suppose we have 3 vectors X_1, X_2 and X_3 that are jointly normally distributed:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \Sigma \right)$$

and that X_2 is independent of X_1 and X_3. We then have no special structure on the means,

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

but the covariance matrix has the form

$$\Sigma = \begin{bmatrix} \Sigma_{11} & 0 & \Sigma_{13} \\ 0 & \Sigma_{22} & 0 \\ \Sigma_{13}^\top & 0 & \Sigma_{33} \end{bmatrix}$$

where $\Sigma_{12} = \Sigma_{21}^\top = \text{cov}(X_1, X_2) = 0$ due to the independence of X_1 and X_2. Similarly for $\Sigma_{23} = \Sigma_{32}^\top = 0$. Because Σ is symmetric, $\Sigma_{31} = \Sigma_{13}^\top$.

Then X_1 and X_2 are conditionally independent given X_3. Written a different way, the two normal random variables, $X_{1|X_3}$ (X_1 conditioned on knowing X_3) and $X_{2|X_3}$ (X_2 conditioned on knowing X_3) are independent.

$\text{Cov}((X_1|X_3), (X_2|X_3)) = 0$

high-level intuition: $X_2 | X_3 \Leftrightarrow X_2$ (bec. X_2 and X_3 are independent)

and $(X_1|X_3)$ are independent.

To see why this is true, we partition Σ as

$$\Sigma = \left[\begin{array}{cc|c} \Sigma_{11} & 0 & \Sigma_{13} \\ 0 & \Sigma_{22} & 0 \\ \hline \Sigma_{13}^\top & 0 & \Sigma_{33} \end{array}\right].$$

We compute the covariance of X_1 and X_2 conditioned on X_3, that is

$$\left[\begin{array}{c} X_1 \\ X_2 \end{array}\right] \Bigg| X_3, \quad = \quad \left[\begin{array}{c} X_1 | X_3 \\ X_2 | X_3 \end{array}\right]$$

using the Schur complement from **Fact 1**

$$\Sigma_{1|2} := \tilde{\Sigma}_{11} - \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_{21}.$$

$$\begin{aligned} \operatorname{cov}\left(\begin{bmatrix} X_{1|X_3} \\ X_{2|X_3} \end{bmatrix}, \begin{bmatrix} X_{1|X_3} \\ X_{2|X_3} \end{bmatrix}\right) &= \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} - \begin{bmatrix} \Sigma_{13} \\ 0 \end{bmatrix} \Sigma_{33}^{-1} \begin{bmatrix} \Sigma_{13}^\top & 0 \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{11} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{13}^\top & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \end{aligned}$$

Because the off-diagonal blocks are zero, the two random variables $X_{1|X_3}$ and $X_{2|X_3}$ are uncorrelated, and because they are normal, we conclude they are independent.

Once again, what we have seen is that if X_1 and X_2 are independent, and we also have X_2 is independent of X_3, then X_1 and X_2 remain independent when we condition them on X_3.

Fact 3: Covariance of a Sum of Independent Normal Random Variables Let X_1 and X_2 be independent normal random vectors, with means μ_1 and μ_2, and covariances, Σ_{11} and Σ_{22}. Define Y as a "linear combination" X_1 and X_2 via

$$Y = AX_1 + BX_2$$

for appropriately sized matrices A and B. Then

$$\mu_Y = A\mu_1 + B\mu_2$$

and

$$\text{cov}(Y, Y) = A\Sigma_{11}A^\top + B\Sigma_{22}B^\top$$

To see why this is true, we first note that

$$(Y - \mu_Y)(Y - \mu_Y)^\top = A(X_1 - \mu_1)(X_1 - \mu_1)^\top A^\top + B(X_2 - \mu_2)(X_2 - \mu_2)^\top B^\top + 2A(X_1 - \mu_1)(X_2 - \mu_2)^\top B^\top$$

And then note that when expectations are taken on each side, the independence of X_1 and X_2 gives

$$\mathcal{E}\{(X_1 - \mu_1)(X_2 - \mu_2)^\top\} = 0.$$

Therefore,

$$\begin{aligned}\text{cov}(Y, Y) &= \mathcal{E}\{(Y - \mu_Y)(Y - \mu_Y)^\top\} \\ &= A\mathcal{E}\{(X_1 - \mu_1)(X_1 - \mu_1)^\top\}A^\top + B\mathcal{E}\{(X_2 - \mu_2)(X_2 - \mu_2)^\top\}B^\top \\ &= A\Sigma_{11}A^\top + B\Sigma_{22}B^\top.\end{aligned}$$

Remark: The next few pages discuss the "information" or "precision" matrix. You will likely encounter it in other courses, such as Mobile Robotics, or in papers. We first saw it when we compared BLUE to Weighted Least Squares (they are the same when the weight chosen as the Information Matrix of the

noise term). You will not need to know anything about the information matrix in the context of the Kalman Filter: when seeing the filter for the first time, you do not need to do every possible variation.

Information or Precision Matrix: The Kalman filter can be written in many forms. One alternative form propagates the inverse of the covariance matrix instead of the covariance matrix. The inverse of the covariance matrix has two names: *information matrix* and *precision matrix*. We will use the first one:

$$\text{Information matrix: } \Lambda := \Sigma^{-1}$$

We decompose it just as we did with the covariance matrix.

$$\Lambda =: \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

The formula for inversion of block matrices gives

$$\begin{aligned}
\Lambda_{11} &= (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \\
\Lambda_{12} &= -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\
\Lambda_{21} &= \Lambda_{12}^{\top} \\
\Lambda_{22} &= \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}
\end{aligned}$$

(See `http://en.wikipedia.org/wiki/Matrix_inversion_lemma#Blockwise_inversion`)

We also scale the mean by defining

$$\eta := \Lambda\mu$$

that is,

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} := \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

Remark: For a multivariate normal distribution, it is equivalent to know η and Λ or μ and Σ. We go back and forth between the two by matrix inversion and multiplication. One sometimes says that these are dual parameterizations for the normal distribution. We only mention the alternative parameterization with the information matrix because sometimes it is easier to use than the more standard mean and covariance representation.

Conditional Distributions Using the Information Matrix: The information matrix of the random variable X_1 given that $X_2 = x_2$ is

$$\Lambda_{1|2} = \Lambda_{11}$$

and

$$\eta_{1|2} = \eta_1 - \Lambda_{12} x_2$$

In other words, if you have the information matrix handy, computing the conditional distribution is easier with it than with the covariance matrix. We note that if you want to go back to the standard representation, then

$$\Sigma_{1|2} = \Lambda_{1|2}^{-1}$$

and

$$\mu_{1|2} = \Lambda_{1|2}^{-1} \, \eta_{1|2}$$

Marginal Distributions Using the Information Matrix: Getting the marginal distributions from the information form of the distribution is more complicated. If you are interested, you can easily find it on the web or in most graduate level probability texts.

Revisiting MVE: for the special case of normal R.V.s (also keep in mind that MVE is a special case of Kalmen filter)

$$y = Cx + \varepsilon, \quad x \in \mathbb{R}^n, \quad x \sim \mathcal{N}(\mu_x, \Sigma_{xx})$$
$$\varepsilon \in \mathbb{R}^m, \quad \varepsilon \sim \mathcal{N}(\mu_\varepsilon, \Sigma_{\varepsilon\varepsilon})$$

$$\mathbb{E}\left\{(x-\mu_x)(\varepsilon-\mu_\varepsilon)^T\right\} = \Sigma_{x\varepsilon} = \Sigma_{\varepsilon x}^T = 0_{n\times m}$$

MVE ($\mu_x = 0$, $\mu_\varepsilon = 0$, $\Sigma_{\varepsilon\varepsilon} = Q$, $\Sigma_{xx} = P$)

$$\hat{x} = PC^T(CPC^T+Q)^{-1}y$$

y ← measurement

By Fact 3:

$$\mu_y = \mathbb{E}(Cx+\varepsilon) = C\mu_x + \mu_\varepsilon$$

$$\Sigma_{yy} = \mathbb{E}\left\{(y-\mu_y)(y-\mu_y)^T\right\}$$

$$= C\Sigma_{xx}C^T + \Sigma_{\varepsilon\varepsilon}$$

$Y = AX_1 + BX_2$

$\operatorname{cov}(Y,Y) = A\Sigma_{11}A^\top + B\Sigma_{22}B^\top$

$Y = CX + I\varepsilon$

$$V = \begin{bmatrix} X \\ Y \end{bmatrix}$$

$$\mathbb{E}(V) = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$$

$$\operatorname{Cov}(V) = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_{yy} \end{bmatrix}$$

$$Cx + \varepsilon - \mu_y$$
$$= Cx + \varepsilon - C\mu_x - \mu_\varepsilon$$

$$\Sigma_{XY} = \mathbb{E}\left\{ (X-\mu_X)(Y-\mu_Y)^T \right\}$$

$$= \mathbb{E}\left\{ (X-\mu_X)(CX+\varepsilon-\mu_Y)^T \right\}$$

$$= \mathbb{E}\left\{ (X-\mu_X)(CX+\varepsilon-C\mu_X-\mu_\varepsilon)^T \right\}$$

$$= \mathbb{E}\left\{ (X-\mu_X)(C(X-\mu_X)+\varepsilon-\mu_\varepsilon)^T \right\}$$

$$= \mathbb{E}\left\{ (X-\mu_X)(X-\mu_X)^T C^T \right\} + \mathbb{E}\left\{ (X-\mu_X)(\varepsilon-\mu_\varepsilon)^T \right\}$$

$$= \Sigma_{XX} C^T + 0$$

$$\mathrm{Cov}(V) = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XX} C^T \\ C\Sigma_{XX} & C\Sigma_{XX} C^T + \Sigma_{\varepsilon\varepsilon} \end{bmatrix}$$

Let's look at $X \mid Y$:

Mean: $\mu_{1|2} := \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$

Covariance: $\Sigma_{1|2} := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

From Fact 1:

$$\mu_{X|Y} = \mu_X + \Sigma_{XX} C^T (C\Sigma_{XX}C^T + \Sigma_{\varepsilon\varepsilon})^{-1}(Y - \mu_Y)$$

$$\Sigma_{X|Y} = \Sigma_{XX} - \Sigma_{XX} C^T (C\Sigma_{XX}C^T + \Sigma_{\varepsilon\varepsilon})^{-1} C\Sigma_{XX}$$

$$\hat{x} = E(X|Y) \quad \text{with} \quad \mu_x = 0, \quad \mu_\varepsilon = 0$$

$$= \mu_{X|Y}$$

$$= PC^T (CPC^T + Q)^{-1} y$$

MVE

$$x_{t+1} = I x_t + G w_t$$

$$y_t = C x_t + I \varepsilon_t$$

MVE ($\mu_x = 0$, $\mu_\varepsilon = 0$, $\Sigma_{\varepsilon\varepsilon} = Q$, $\Sigma_{xx} = P$)

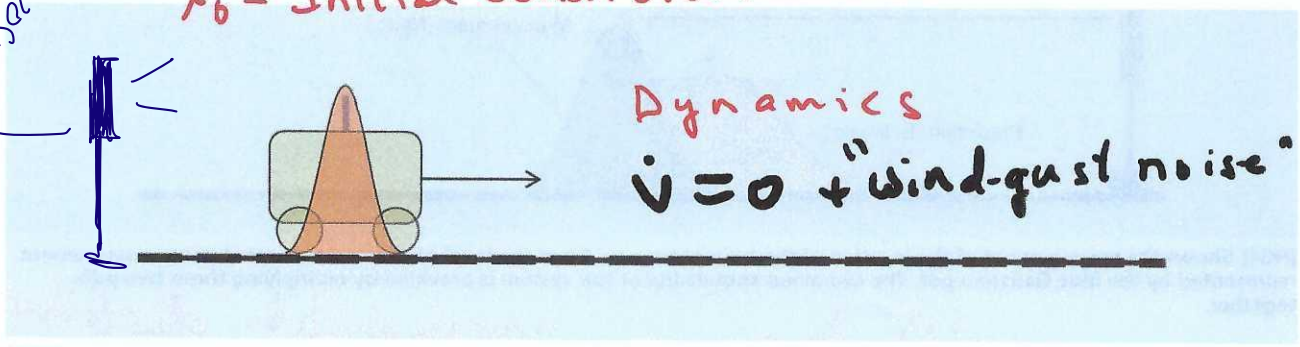
$\hat{x} = PC^T (CPC^T + Q)^{-1} y$ ← measurement

BX

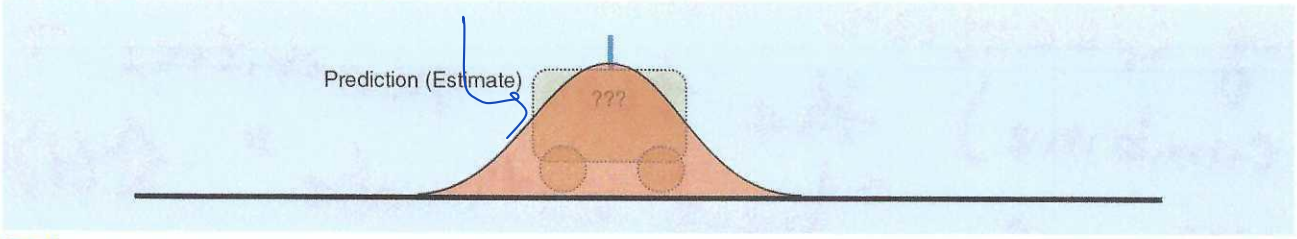
Kalman Filter Motivation

X_0 = Initial Condition

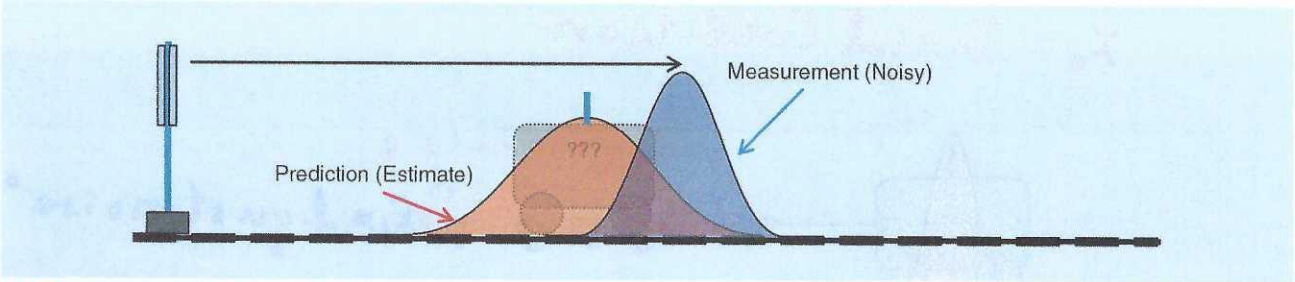
range sensor



Uncertainty typically grows as time increases



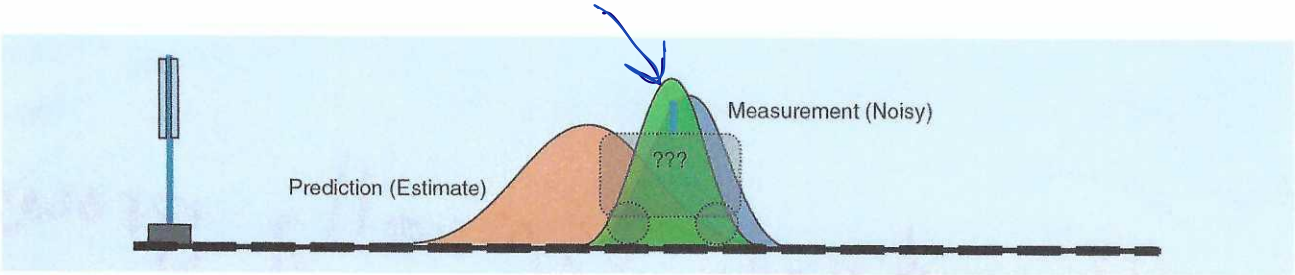
[FIG3] Here, the prediction of the location of the train at time $t = 1$ and the level of uncertainty in that prediction is shown. The confidence in the knowledge of the position of the train has decreased, as we are not certain if the train has undergone any accelerations or decelerations in the intervening period from $t = 0$ to $t = 1$.



[FIG4] Shows the measurement of the location of the train at time $t = 1$ and the level of uncertainty in that noisy measurement, represented by the blue Gaussian pdf. The combined knowledge of this system is provided by multiplying these two pdfs together.

$x(t)$ has a "probability distribution"

$y(t)$ is not perfect = "probability distribution"



[FIG5] Shows the new pdf (green) generated by multiplying the pdfs associated with the prediction and measurement of the train's location at time $t = 1$. This new pdf provides the best estimate of the location of the train, by fusing the data from the prediction and the measurement.

Big Question: How to "fuse" (= combine) the two quantities to have a "best estimate" $\hat{x}(t)$???

RLS: $A_k = I$, $G_k = 0$

V_k = deterministic = = no stochastic or random model.

Rob 501 Handout: Grizzle
Discrete-Time Kalman Filter and its Derivation

Model Linear time-varying discrete-time system with "white[1]" Gaussian noise

$$\begin{aligned} x_{k+1} &= A_k x_k + G_k w_k, \quad x_0 \text{ initial condition} \\ y_k &= C_k x_k + v_k \end{aligned}$$

$x \in \mathbb{R}^n$, $w \in \mathbb{R}^p$, $y \in \mathbb{R}^m$, $v \in \mathbb{R}^m$. Moreover, the random vectors x_0, and, for $k \geq 0$, w_k, v_k are all independent[2] Gaussian (normal) random vectors.

Notation: $\delta_{kl} = 1 \Leftrightarrow k = l$ (and $\delta_{kl} = 0, \quad k \neq l$)

Precise Assumptions on the Random Variables

- For all $k \geq 0,\ l \geq 0$, x_0, w_k, v_l are jointly Gaussian.
- w_k is a 0-mean white noise process: $\mathcal{E}\{w_k\} = 0$, and $\text{cov}\,(w_k, w_l) = R_k \delta_{kl}$
- v_k is a 0-mean white noise process: $\mathcal{E}\{v_k\} = 0$, and $\text{cov}\,(v_k, v_l) = Q_k \delta_{kl}$
- Uncorrelated noise processes: $\text{cov}\,(w_k, v_l) = 0$
- The initial condition x_0 is uncorrelated with all other noise sequences.
- We denote the mean and covariance of x_0 by

$$\bar{x}_0 = \mathcal{E}\{x_0\} \quad \text{and} \quad P_0 = \text{cov}(x_0) = \text{cov}\,(x_0, x_0) = \mathcal{E}\{(x_0 - \bar{x}_0)\,(x_0 - \bar{x}_0)^\top\}$$

[1] Recall that in white light, all frequencies are present. When only certain frequency components are present, you get "colored" light, such as blue light or red light. The term "white" noise means that if you compute the power spectral density of the noise random process, it is a constant, meaning that all frequency components are equally represented, just as in white light.

[2] Recall that for normal random variables, uncorrelated and independent are the same thing. This is one of several special properties of Gaussian random variables.

Short-hand notation for the noise modeling assumptions:

$$\text{cov}\left(\begin{bmatrix} w_k \\ v_k \\ x_0 \end{bmatrix}, \begin{bmatrix} w_l \\ v_l \\ x_0 \end{bmatrix}\right) = \begin{bmatrix} R_k\delta_{kl} & 0 & 0 \\ 0 & Q_k\delta_{kl} & 0 \\ 0 & 0 & P_0 \end{bmatrix}, \quad \delta_{kl} = \begin{bmatrix} 1, & k = l \\ 0, & k \neq l \end{bmatrix}$$

Lemma (Propertis of x_k and y_k Coming from the Model)

- For all $k \geq 1$, x_k is a linear combination of x_0 and $w_0, \cdots, w_{k-1}$. In particular, x_k is uncorrelated with w_k.
- For all $k \geq 1$, y_k is a linear combination of x_0, $w_0, \cdots, w_{k-1}$, and $v_0, \cdots, v_k$. In particular, y_k is uncorrelated with w_k .
- For all $k \geq 0$, v_k is uncorrelated with x_k.

The proof is by induction using the recursive nature of the discrete-time model. We skip it. The reader can easily fill it in.

Remark: On the next page, we give (one form of) the discrete-time Kalman Filter. After that, we do the main elements of its derivation. There are many variations of the basic filter, all equivalent to the one we give, but some preferable over others for numerical reasons. Page 8 provides a version of the filter with the measurement update and prediction steps combined.

The Kalman Filter

Definition of Terms:

$$\widehat{x}_{k|k} := \mathcal{E}\{x_k | y_0, \cdots, y_k\}$$
$$P_{k|k} := \mathcal{E}\{(x_k - \widehat{x}_{k|k})(x_k - \widehat{x}_{k|k})^\top | y_0, \cdots, y_k\}$$

$$\widehat{x}_{k+1|k} := \mathcal{E}\{x_{k+1} | y_0, \cdots, y_k\}$$
$$P_{k+1|k} := \mathcal{E}\{(x_{k+1} - \widehat{x}_{k+1|k})(x_{k+1} - \widehat{x}_{k+1|k})^\top | y_0, \cdots, y_k\}$$

Initial Conditions:

$$\widehat{x}_{0|-1} := \bar{x}_0 = \mathcal{E}\{x_0\}, \quad \text{and} \quad P_{0|-1} := P_0 = \text{cov}(x_0)$$

For $k \geq 0$

Measurement Update Step:

$$K_k = P_{k|k-1} C_k^\top \left(C_k P_{k|k-1} C_k^\top + Q_k \right)^{-1}$$
(Kalman Gain)
$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + K_k \left(y_k - C_k \widehat{x}_{k|k-1} \right)$$
$$P_{k|k} = P_{k|k-1} - K_k C_k P_{k|k-1}$$

Time Update or Prediction Step:

$$\widehat{x}_{k+1|k} = A_k \widehat{x}_{k|k}$$
$$P_{k+1|k} = A_k P_{k|k} A_k^\top + G_k R_k G_k^\top$$

End of For Loop (Just stated this way to emphasize the recursive nature of the filter)

Preliminaries

Measurements: We collect all of the measurements at time k as

$$Y_k = (y_k, y_{k-1}, \cdots, y_0).$$

Strictly speaking, we should be stacking them up into a column vector as we have done for all of our estimation problems, but notationally, it is more convenient to write them in a row. Also, it is more convenient to put the most recent measurement at the head of the list. We note that

$$Y_k = (y_k, Y_{k-1}).$$

Hence,

$$\begin{aligned}
\widehat{x}_{k|k} :=& \mathcal{E}\{x_k|Y_k\} \\
P_{k|k} :=& \mathcal{E}\{(x_k - \widehat{x}_{k|k})(x_k - \widehat{x}_{k|k})^\top|Y_k\} \\
& \text{mean and covariance of the conditional normal random vector } \; x_k|Y_k
\end{aligned}$$

$$\begin{aligned}
\widehat{x}_{k+1|k} :=& \mathcal{E}\{x_{k+1}|Y_k\} \\
P_{k+1|k} :=& \mathcal{E}\{(x_{k+1} - \widehat{x}_{k+1|k})(x_{k+1} - \widehat{x}_{k+1|k})^\top|Y_k\} \\
& \text{mean and covariance of the conditional normal random vector } \; x_{k+1}|Y_k
\end{aligned}$$

Important Remarks:

- The conditional random vector $x_k|Y_k$ is distributed $N(\widehat{x}_{k|k}, P_{k|k})$.
- The conditional random vector $x_{k+1}|Y_k$ is distributed $N(\widehat{x}_{k+1|k}, P_{k+1|k})$.

Filter Derivation Using Induction and Properties of Conditional Distributions of Gaussian Random Vectors

Base step: The initial conditions of the filter at time $k = 0$, namely

$$\widehat{x}_{0|-1} := \bar{x}_0, \quad \text{and} \quad P_{0|-1} := P_0$$

Induction step: At time $k \geq 0$, we suppose that $(\widehat{x}_{k|k-1}, P_{k|k-1})$ are known, and we derive $(\widehat{x}_{k|k}, P_{k|k})$ and $(\widehat{x}_{k+1|k}, P_{k+1|k})$.

Key idea of the development: We need to compute the distribution (or density) of the conditional random vector

$$x_k|Y_k = x_k|(y_k, Y_{k-1})$$

From HW 7, we learned **Fact 4** $X|(Y,Z) = X|Z \;\Big|\; Y|Z$. From this we obtain

$$x_k|Y_k = x_k|(y_k, Y_{k-1}) = x_k|Y_{k-1} \;\Big|\; y_k|Y_{k-1} \quad (*)$$

where we have identified

$$x_k \leftrightarrow X, \quad y_k \leftrightarrow Y, \quad \text{and} \quad Y_{k-1} \leftrightarrow Z.$$

Hence, if we can compute the distribution (or density) of

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} \Big| Y_{k-1}$$

then we can apply **Fact 1** to obtain (*).

The following calculations are aimed at doing just this.

Measurement Update: We seek to derive the filter equations on page 3. To begin, we have that $y_k = C_k x_k + v_k$. It follows by linearity that the conditional random variable $y_k|Y_{k-1}$ is equal to

$$y_k|Y_{k-1} = C_k \; x_k|Y_{k-1} + v_k|Y_{k-1}.$$

From our assumptions on the noise, v_k is independent of both x_k and Y_{k-1}, and hence by **Fact 2**, $v_k|Y_{k-1}$ is independent of the conditional random variable $x_k|Y_{k-1}$. Moreover, because v_k is independent of Y_{k-1}, $v_k|Y_{k-1} = v_k$. Putting this together, we have that

$$y_k|Y_{k-1} = C_k \; x_k|Y_{k-1} + v_k,$$

and $x_k|Y_{k-1}$ and v_k are independent. Hence

$$\begin{aligned}
\widehat{y}_{k|k-1} :=& \mathcal{E}\{y_k|Y_{k-1}\} \\
=& \mathcal{E}\{C_k x_k|Y_{k-1}\} + \mathcal{E}\{v_k\} \\
=& C_k \mathcal{E}\{x_k|Y_{k-1}\} + \mathcal{E}\{v_k\} \\
=& C_k \widehat{x}_{k|k-1} + 0 \\
=& C_k \widehat{x}_{k|k-1}.
\end{aligned}$$

Moreover, the independence of $x_k|Y_{k-1}$ and v_k with **Fact 3** yields

$$\mathrm{cov}(y_k|Y_{k-1}, y_k|Y_{k-1}) = C_k P_{k|k-1} C_k^\top + Q_k.$$

We use independence again to obtain

$$\mathrm{cov}(x_k|Y_{k-1}, y_k|Y_{k-1}) = \mathrm{cov}(x_k|Y_{k-1}, C_k \; x_k|Y_{k-1}) = P_{k|k-1} C_k^\top.$$

With this information, we conclude that the vector

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} | Y_{k-1}$$

is jointly normally distributed, with mean and covariance

$$\begin{bmatrix} \widehat{x}_{k|k-1} \\ C_k \widehat{x}_{k|k-1} \end{bmatrix}, \begin{bmatrix} P_{k|k-1} & P_{k|k-1} C_k^\top \\ C_k P_{k|k-1} & C_k P_{k|k-1} C_k^\top + Q_k \end{bmatrix} \qquad (*)$$

As discussed on the previous page, to compute the distribution of $(x_k|Y_k)$, we have from **Fact 4**

$$x_k|Y_k = x_k \Big| (y_k, Y_{k-1}) = x_k|Y_{k-1} \;\Big|\; y_k|Y_{k-1},$$

and thus applying **Fact 1** to **(*)** we compute the mean and covariance of $x_k|Y_k = x_k|Y_{k-1} \;\Big|\; y_k|Y_{k-1}$ to be

$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + P_{k|k-1} C_k^\top \left[C_k P_{k|k-1} C_k^\top + Q_k \right]^{-1} \left(y_k - C_k \widehat{x}_{k|k-1} \right)$$

$$P_{k|k} = P_{k|k-1} - P_{k|k-1} C_k^\top \left[C_k P_{k|k-1} C_k^\top + Q_k \right]^{-1} C_k P_{k|k-1}$$

Remark: We note that $P_{k|k}$ is the Schur complement $C_k P_{k|k-1} C_k^\top + Q_k$ in the covariance of

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} | Y_{k-1}$$

Prediction or Time Update: We seek to derive the filter equations on page 3. This time we use the state-variable model instead of the output model, namely

$$x_{k+1} = A_k x_k + G_k w_k,$$

and we are interested in the random vector

$$x_{k+1}|Y_k = A_k x_k|Y_k + G_k w_k|Y_k.$$

Because x_k and Y_k are both independent of w_k, by **Fact 2**, $x_k|Y_k$ and $w_k|Y_k$ are also independent. It follows that

$$\begin{aligned}
\widehat{x}_{k+1|k} &= \mathcal{E}\{x_{k+1}|Y_k\} \\
&= \mathcal{E}\{A_k x_k + G_k w_k|Y_k\} \\
&= A_k \mathcal{E}\{x_k|Y_k\} + G_k \mathcal{E}\{w_k|Y_k\} \\
&= A_k \widehat{x}_{k|k} + G_k \mathcal{E}\{w_k\} \\
&= A_k \widehat{x}_{k|k},
\end{aligned}$$

where we have used $w_k|Y_k = w_k$, and $\mathcal{E}\{w_k\} = 0$.

Next, we use **Fact 3** and the conditional independence of the random vectors $x_k|Y_k$ and $w_k|Y_k$ to evaluate the covariance of $x_{k+1}|Y_k$ as

$$P_{k+1|k} = A_k P_{k|k} A_k^\top + G_k R_k G_k^\top.$$

That's the Proof Folks!

Combined Update Version of the Filter (often easier to implement)

Here, we will assume the model also has a *deterministic* input u_k, and thus

$$x_{k+1} = A_k x_k + B_k u_k + G_k w_k$$
$$y_k = C_k x_k + v_k,$$

with the assumptions on the random vectors x_0, w_k and v_k the same as before.

Combined Filter: The measurement-update step and time-update step of the Kalman Filter can be combined into a single step. The algorithm becomes:

Initial Conditions:

$$\widehat{x}_{0|-1} := \bar{x}_0 = \mathcal{E}\{x_0\}, \quad \text{and} \quad P_{0|-1} := P_0 = \text{cov}(x_0)$$

For $k \geq 0$

$$K_k = (P_{k|k-1} C_k^\top) \left[C_k P_{k|k-1} C_k^\top + Q_k \right]^{-1}$$

$$\widehat{x}_{k+1|k} = A_k \widehat{x}_{k|k-1} + B_k u_k + A_k K_k \left(y_k - C_k \widehat{x}_{k|k-1} \right)$$

$$P_{k+1|k} = A_k \left[P_{k|k-1} - K_k C_k P_{k|k-1} \right] A_k^\top + G_k R_k G_k^\top$$

End of For Loop

Remark: You do not have to start at $k = 0$. In MATLAB, it is often easier to begin with $k = 1$. In that case, the initial conditions are

$$\widehat{x}_{1|0} := \bar{x}_0 = \mathcal{E}\{x_0\}, \quad \text{and} \quad P_{1|0} := P_0 = \text{cov}(x_0)$$

Remark: $K_k C_k P_{k|k-1} = (P_{k|k-1} C_k^\top) \left[C_k P_{k|k-1} C_k^\top + Q_k \right]^{-1} C_k P_{k|k-1}$ is symmetric positive semi-definite and represents the value of the measurement in reducing the covariance of the state estimate, just as in the MVE.