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: Jointly 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

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

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

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

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

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

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

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:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

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.

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

$$\left[\begin{matrix} X_1 \\ X_2 \end{matrix}\right] \Bigg| X_3,$$

using the Schur complement from **Fact 1**

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

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

$$\begin{aligned}(Y - \mu_Y)(Y - \mu_Y)^\top = A(X_1 - \mu_1)(X_1 - \mu_1)^\top A^\top + B(X_2 - \mu_2)(X_2 - \mu_2)^\top B^\top \\ + 2A(X_1 - \mu_1)(X_2 - \mu_2)^\top B^\top\end{aligned}$$

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.