Indian Institute of Technology Bombay
Department of Electrical Engineering

Handout 11 EE 325 Probability and Random Processes
Lecture Notes 12 March 27, 2019

1 Continuous Valued Random Variables

Once we thoroughly understand discrete random variables, a conceptual extension to continuous valued random variables is straightforward. However, the computations here are often plagued by convergence issues of the integral. There are excellent theories and convergence concepts(measure theory) which will assure us a safe-passage through this apparent difficulties. However we do not resort to such approaches, rather we merely state some pre-requisites or conditions such that the quantities we compute are well-defined.

In our discussions, the properties of continuous valued random variables will appear similar to that of discrete valued RVs. The major difference is that while we had the luxury to consider countable sets of the form $\{X = x\}$ for discrete RVs, we have to look at sets which generate the Borel sigma-field or $\mathcal{B}(\mathbb{R})$ in the continuous counterparts. We already learned that $\mathcal{B}(\mathbb{R})$ is generated by the open-sets, and on $\mathbb{R}$ we can consider semi-infinite intervals of the form $(-\infty, x]$. An appropriate change of the so called seed-sets from $\{X = x\}$ in the discrete case to $\{X \leq x\}$ for the continuous case, allows us to reuse most concepts which were already learnt. In fact, this is a readymade formula for a discrete to real converter for probability theory.

Definition 1 *A function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a* ***random variable*** *if $\forall a \in \mathbb{R}$,*

$$\{X \leq a\} \in \mathcal{F},$$

and $(\Omega, \mathcal{F}, P)$ is a probability space.

Notice that we did not insist in calling X a continuous valued random variable. The reason is that even when X is discrete, our definition guarantees that $\{X = x\}, x \in E_1$ will be in $\mathcal{F}$. Thus our definition covers the discrete case with X taking values in a countable set E_1, as well as the real-valued case. This statement should set the tone for our discussion here, we will use the word *random variable* to represent the general case. Recall the meaning

$$\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\}.$$

Thus $\{X \leq x\} \in \mathcal{F}$ will imply that we will be able to assign a probability to every such element, as the measure $P(\cdot)$ is defined for all elements of $\mathcal{F}$. In this sense, $\{(-\infty, x]\}$ are the seed-sets that we discussed previously, with $\sigma((-\infty, x]) = \mathcal{B}(\mathbb{R})$.

Definition 2 *The cumulative distribution function (cdf) is the induced probability measure defined as*

$$P\left((\infty, x]\right) = P(X \leq x).$$

Clearly, since the cdf defines probability for the class of sets which generate $\mathcal{B}(\mathbb{R})$, we can extend the cdf to $\mathcal{B}(\mathbb{R})$. That such an extension exists, and is unique, follows from the properties that we discussed in earlier lectures. We will not revisit these at this point, rather assume that it is sufficient to consider $\{X \leq x\}$. The cdf of X is usually denoted as $F_X(x)$.

Theorem 1 *The cdf $F_X(x)$ is right continuous and monotone non-decreasing. Furthermore, if X is $\mathbb{R}-$valued*

$$F(\infty) = \lim_{x\uparrow\infty} F(x) = 1 \text{ and } F(-\infty) = \lim_{x\downarrow-\infty} F(x) = 0.$$

Proof: The proof is left as an exercise from class notes. While the cdf $F_X(x)$ of a random variable completely specifies it, many a times it is convenient to work with the so called probability density functions. However, many random variables may not have such a representation, for example, when the random variable also takes a few discrete values. Discrete random variables have a mass distribution function as opposed to a density function.

Definition 3 *Consider a non-negative function $f(\cdot)$ such that*

$$\int_{\mathbb{R}} f(x)dx = 1.$$

If the cdf $F_X(x)$ of a $\mathbb{R}-$ valued random variable can be represented as

$$F_X(x) = \int_{-\infty}^{x} f(u)du, \forall x \in \mathbb{R},$$

then X is said to admit a density $f(x)$. In this case, the random variable X is also called absolutely continuous (or absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$).

One standard trick that we will repeatedly perform is to write the cdf of any random variable in the integral form shown above. Then, we can compute the density function, if it exists, by mere identification. The identification will give us a pdf which is unique, except at a set of measure zero. This later condition is due to the fact that two functions differing only at a countable set of real values is equivalent under integral operations.

General random variables that we deal can possibly have a discontinuous cdf. We can deal with the discontinuities separately by considering the cdf as the summation of two parts. One part corresponding to a discrete random variable and another part for the absolutely continuous component. Let us use this idea to define integrals in a convenient way.

Definition 4 *Consider a cdf $F_X(x)$ with a countable number of discontinuities occurring at $\{d_n\}, n \in \mathbb{N}$. The integral w.r.t a probability measure is defined as,*

$$\int_{-\infty}^{+\infty} g(x)dF_X(x) = \sum_{n\in N} g(d_n)\left(F_X(d_n) - F_X(d_{n-})\right) + \int_{-\infty}^{+\infty} g(x)f_c(x)dx \qquad (1)$$

where $f_c(x)$ is the absolutely continuous component of X.

That we can separate the continuous and discrete parts is illustrated in Figure 1, where d_1 and d_2 are the discontinuities. The function is right continuous, notice the way $F_x(d_1)$ is marked in the figure. The integral with respect to this cdf can be broken into two parts. One corresponding to the discrete part, shown in Figure 2 as F_X^d, and a continuous and differentiable component shown in Figure 3 as F_X^c. It is evident that the integral with respect to $F_X(x)$ can be computed separately for the two parts and superposed later using linearity. The respective outputs in this example are

$$\int g(x)dF_X^d(x) = g(d_1)(F(d_1) - F(d_1^-) + g(d_2)(F_(d_2) - F_(d_2^-)).$$

$$\int g(x)dF_X^c(x) = \int g(x)f_c(x)dx,$$

where $f_c(x)$ is the density function of $F_X^c(x)$. For those are familiar with the theory of distributions in Fourier Transform etc, F_X^d corresponds to a density with two impulses of appropriate heights at d_1 and d_2.

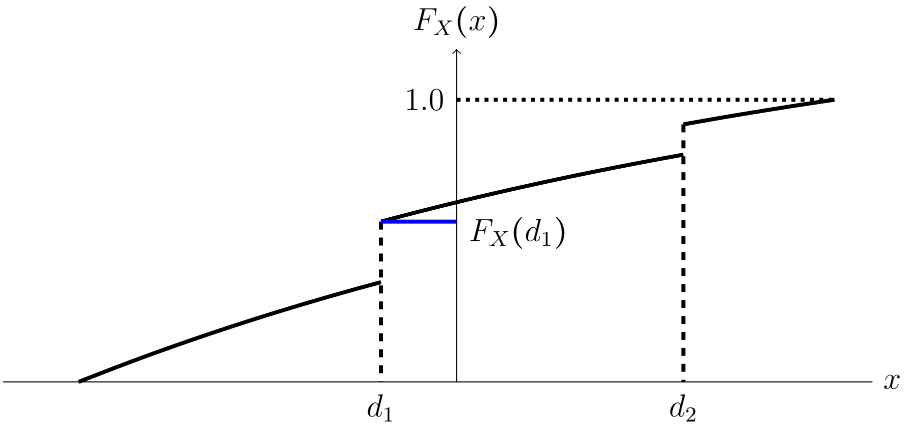

Figure 1: A cdf with discontinuities

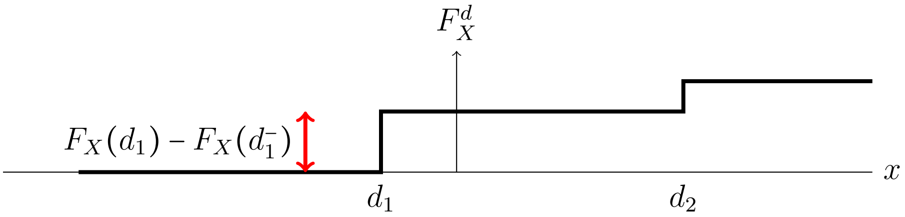

Figure 2: Discrete part of the cdf

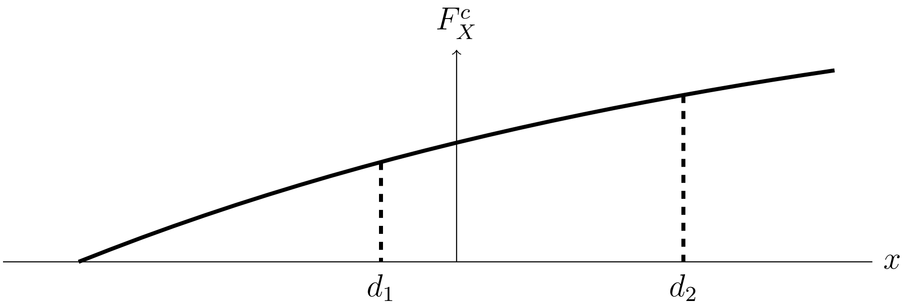

Figure 3: Differentiable part of the cdf

2 Random Vectors

Like in the discrete case, a collection of random variables is called a random vector. Let us first consider two random variables X_1 and X_2.

Definition 5 *A vector* $(X_1, X_2) : \Omega_1 \times \Omega_2 \to \mathbb{R} \times \mathbb{R}$ *is a random vector if* $\{X_1 \leq x_1, X_2 \leq x_2\} \in \mathcal{F}_1 \times \mathcal{F}_2$, $\forall (x_1, x_2)$.

Recall the definition of the product sigma-field $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$, which is the smallest sigma-field containing events of the form $A_1 \times A_2$ with $A_i \in \mathcal{F}_i, i = 1, 2$. For example, we already demonstrated the construction of a probability measure in $\mathbb{R}^2$, by considering rectangles as the seed-sets (refer chapters 2-3).

Consider a probability measure $P(A \times B)$ defined on $\mathcal{F}$. The random vector (X_1, X_2) induces a probability on $(\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R}))$. Thus it makes sense to talk about joint events generated by (X_1, X_2). We can easily generalize the above definition to $\mathbb{R}^n$.

2.1 Probability Density

The pdf of random vectors can be defined akin to random variables.

Definition 6 *Let* $f : \mathbb{R}^n \to R^+$ *be a function such that*

$$\int_{R^n} f(x_1, \cdots x_n) \, dx_1 \cdots dx_n = 1.$$

A random vector $\bar{X} = (X_1, \cdots, X_n)$ *admits a density* $f(\cdot)$ *if*

$$P(\bar{X} \in C) = \int_{\bar{x} \in C} f(x_1, \cdots x_n) dx_1 \cdots dx_n.$$

When the random vector $\bar{X}$ admits a density, it is said to be absolutely continuous.

2.2 Independence

The notion of independence of RVs translates to the independence of events associated with X_1 and X_2. The relevant events can be expressed as $\{X_i \leq x_i\}, i = 1, 2$.

Definition 7 *Random Variables* X_1 *and* X_2 *are independent if*

$$P(X_1 \leq x_1, X_2 \leq x_2) = P(X_1 \leq x_1) P(X_2 \leq x_2).$$

3 Expectation

Many properties of the expectation from the discrete random variable case carries over to the continuous valued ones. In particular, expectation in the former is a weighted summation, where $g(x)$ is weighted by $P(X = x)$ and summed to obtain $\mathbb{E}[g(X)]$. In a more general form, $g(x)$ will be integrated with respect to the measure $dF_X(x)$ to obtain $\mathbb{E}[g(X)]$.

Definition 8 *For a function* $g(\cdot)$ *which is either non-negative or integrable*

$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(x) dF_X(x) \tag{2}$$

Notice that if the absolutely continuous component of X is identically zero, the expectation reduces to the summation as in the discrete case, see (1). Let us look at another intuitive result which also has an exact counterpart in the discrete case.

Theorem 2 *Let X admit a density $f_X(x)$. Then,*

$$\mathbb{E}\mathbb{1}_{\{X\in A\}} = P(A).$$

Proof: By definition,

$$\begin{aligned}\mathbb{E}\mathbb{1}_{\{X\in A\}} &= \int \mathbb{1}_{\{x\in A\}} f_X(x)dx\\ &= \int_A f_X(x)dx\\ &= P(A).\end{aligned}$$

Thus probabilities can be written as expectation. This is very useful, since many properties of expectation are not only easier to derive, but also carry over from the discrete case with appropriate change to integrals. For example, we know that when X and Y are independent discrete RVs, $\mathbb{E}[g_1(X)g_2(Y)] = \mathbb{E}[g_1(X)]\mathbb{E}[g_2(Y)]$. The same holds true for the continuous case also. The key here is the decomposition of joint density into a product form.

Theorem 3 *If X_1 and X_2 are independent random variables admitting respective densities,*

$$f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2).$$

Proof: By the definition of independence, the events $\{X_1 \le x_1\}$ and $\{X_2 \le x_2\}$ are independent.

$$\begin{aligned}P(X_1 \le x_1, X_2 \le x_2) &= P(X_1 \le x_1)P(X_2 \le x_2)\\ &= \int_{-\infty}^{x_1} f_{X_1}(u)du \int_{-\infty}^{x_2} f_{X_2}(v)dv\\ &= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f_{X_1}(u)f_{X_2}(v)dv\end{aligned}$$

By identification, the joint pdf is

$$f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2).$$

Theorem 4 *For independent random variables X_1 and X_2*

$$\mathbb{E}g_1(X_1)g_2(X_2) = \mathbb{E}g_1(X_1)\mathbb{E}g_2(X_2).$$

Proof: Notice that we didn't explicitly state the assumption of existence, it is implied when we use the expectation symbol.

$$\begin{aligned}\mathbb{E}g_1(X_1)g_2(X_2) &= \int\int g_1(x_1)g_2(x_2)f_{X_1,X_2}(x_1,x_2)dx_1dx_2\\ &= \int\int g_1(x_1)g_2(x_2)f_{X_1}(x_1)f_{X_2}(x_2)dx_1dx_2\\ &= \int g_1(x_1)f_{X_1}(x_1)dx_1 \int g_2(x_2)f_{X_2}(x_2)dx_2\\ &= \mathbb{E}g_1(X_1)\mathbb{E}g_2(X_2).\end{aligned}$$

The above result can be extended to case where the product $g_1(\cdot)g_2(\cdot)$ is replaced by a joint form $g(\cdot,\cdot)$.

Theorem 5 *Let X_1, X_2 be independent random variables. Consider a function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, which is either non-negative or integrable. Then*

$$\mathbb{E}g(X_1, X_2) = \int f_{X_1}(x_1)\mathbb{E}g(x_1, X_2)dx_1.$$

Proof:

$$\begin{aligned}\mathbb{E}g(X_1, X_2) &= \int \int f_{X_1}(x_1) f_{X_2}(x_2) g(x_1, x_2) dx_1 dx_2 \\ &= \int f_{X_1}(x_1) \left(\int f_{X_2}(x_2) g(x_1, x_2) dx_2 \right) dx_1 \\ &= \int f_{X_1}(x_1)\mathbb{E}g(x_1, X_2)dx_1.\end{aligned}$$

The last theorem will find several applications, some of them we will illustrate in the coming sections.

4 Gaussian Random Vectors

We have introduced the basic properties of Gaussian random variables. In particular, it is an absolutely continuous random variable with the density

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

where $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{E}[(X-\mu)^2]$. Perhaps, this is the single most important distribution in probability theory, and our discussions will get considerably simpler by familiarizing the basic properties of Gaussian distribution. Gaussian distribution is denoted a $\mathcal{N}(\mu, \sigma^2)$. As the notation shows, the distribution is parametrized by two real values, the mean μ and variance σ^2.

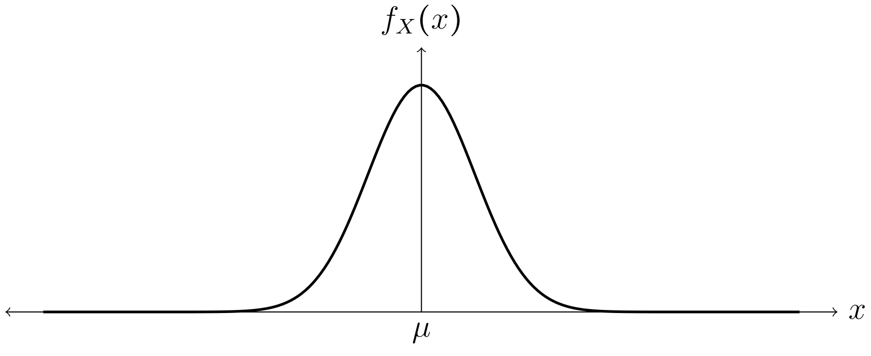

Figure 4: Gaussian Distribution in $\mathbb{R}$

Exercise 1 *Find the value of x for which a zero mean Gaussian distribution packs 99% of the probability in the interval $[-x, x]$.*

Exercise 2 *Compute the cdf of the Gaussian random variable.*

Let us now find the density of Gaussian vectors.

Theorem 6 *If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent Gaussian random variables, the joint density function of (X_1, X_2) is given by,*

$$f(X_1, X_2)(x_1, x_2) = \frac{1}{\sqrt{\det(2\pi K)}} e^{-\frac{1}{2}\begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{bmatrix} K^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}}, \tag{3}$$

where

$$K = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}. \tag{4}$$

Proof: By the product decomposition of independent densities,

$$\begin{aligned} f_{X_1,X_2}(x_1, x_2) &= f_{X_1}(x_1) f_{X_2} x_2 \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} . \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} . \end{aligned}$$

For the diagonal K, it is easy to show that

$$\begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{bmatrix} K^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} = \frac{(x - \mu_1)^2}{2\sigma_1^2} + \frac{(x - \mu_2)^2}{2\sigma_2^2}.$$

■

Let us now find the joint distribution of Gaussians which are not independent. What does it mean by 'Gaussians which are not independent'? One way to think about this is to start with two independent Gaussians and take linear combinations of it. The linear combinations of independent Gaussians are also Gaussians, but such combinations are not necessarily dependent. We have not proved this statement. While this can be done by basic probability tools, let us wait till we introduce the more elegant generating function framework, and proceed now by taking linear combinations.

Theorem 7 *Let X and Z be two independent and identical Gaussians with $X \sim \mathcal{N}(0, \sigma^2)$. Consider the random vector (X_1, X_2) as*

$$X_1 = X\ ;\ X_2 = \sqrt{a}X + \sqrt{1-a}Z.$$

Then with $\bar{x} = (x_1, x_2)^T$,

$$f_{X_1,X_2}(\bar{x}) = \frac{1}{\sqrt{\det(2\pi K)}} e^{-\frac{1}{2}x^T K^{-1} x},$$

where

$$K == \sigma^2 \begin{bmatrix} 1 & \sqrt{a} \\ \sqrt{a} & 1 \end{bmatrix}. \tag{5}$$

Proof: We have used the scale $\alpha = \sqrt{a}$ and $\beta = \sqrt{1-a}$ so that X_1 and X_2 have the same variance. Let us compute pdf of (X_1, X_2). We start with the joint cdf

$$P(X_1 \le x_1, X_2 \le x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{X_1 X_2}(u, v) du dv. \tag{6}$$

Once we get the form in the RHS, we can obtain the joint pdf by identification.

$$\begin{aligned}
P(X_1 \le x_1, X_2 \le x_2) &= \mathbb{E}\mathbb{1}_{\{X_1 \le x_1, X_2 \le x_2\}} \\
&= \mathbb{E}\mathbb{1}_{\{X_1 \le x_1\}}\mathbb{1}_{\{X_2 \le x_2\}} \\
&= \mathbb{E}\mathbb{1}_{\{X \le x_1\}}\mathbb{1}_{\{\alpha X + \beta Z \le x_2\}} \\
&= \int_{-\infty}^{x_1}\int_{-\infty}^{\frac{1}{\beta}(x_2-\alpha u)} f_{X,Z}(u,v)dudv \\
&= \int_{-\infty}^{x_1}\int_{-\infty}^{\frac{1}{\beta}(x_2-\alpha u)} f_X(u)f_Z(v)dudv \\
&= \int_{-\infty}^{x_1} f_X(u)\left(\int_{-\infty}^{\frac{1}{\beta}(x_2-\alpha u)} f_Z(v)dv\right)du
\end{aligned}$$

By the change of variables $w = \beta v + \alpha u$,

$$P(X_1 \le x_1, X_2 \le x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2}\frac{1}{\beta}f_X(u)f_Z(\frac{w-\alpha u}{\beta})dwdu.$$

From this, the joint pdf of (X_1, X_2) is

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{\beta}f_X(x_1)f_Z(\frac{x_2-\alpha x_1}{\beta}).$$

Observe that our derivation is valid for any X_1, X_2 such that $X_2 = \alpha X_1 + \beta Z$, where Z is independent of X_1. Let us now specialize our result to the Gaussian setting.

$$\begin{aligned}
\frac{1}{\beta}f_X(x_1)f_Z(\frac{x_2-\alpha x_1}{\beta}) &= \frac{1}{\beta}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x_1^2}{2\sigma^2}}.\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_2-\alpha x_1)^2}{2\beta^2\sigma^2}} \\
&= \frac{1}{2\pi\sigma^2\beta}e^{-\frac{x_1^2}{2\sigma^2}(1+\frac{\alpha^2}{\beta^2})}e^{-\frac{x_2^2}{2\beta^2\sigma^2}}e^{2\frac{\alpha x_1 x_2}{2\beta^2\sigma^2}} \\
&= \frac{1}{2\pi\sigma^2\beta}e^{-\frac{x_1^2}{2\beta^2\sigma^2}}e^{-\frac{x_2^2}{2\beta^2\sigma^2}}e^{2\frac{\alpha x_1 x_2}{2\beta^2\sigma^2}} \\
&= \frac{1}{\sqrt{(\det 2\pi K)}}e^{-\bar{x}^T K^{-1}\bar{x}},. \qquad (7)
\end{aligned}$$

The matrix K has the form in (5) because we have assumed zero mean identical Gaussians. However, it can be easily generalized to arbitrary Gaussian vectors. While similar to the one we introduced in (4), the matrix in this question is no more diagonal. Whether the matrix K is diagonal or not is a strong indication of the independence of the participating variables. We can easily invert the matrix K to obtain

$$K^{-1} = \frac{1}{\sigma^2(1-\alpha^2)}\begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix} = \frac{1}{\sigma^2\beta^2}\begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix} \qquad (8)$$

Also,

$$\det 2\pi K = \det\begin{bmatrix} 2\pi\sigma^2 & \alpha 2\pi\sigma^2 \\ \alpha 2\pi\sigma^2 & 2\pi\sigma^2 \end{bmatrix} = 4\pi^2\sigma^4\beta^2.$$

The elegant form of the Gaussian pdf allows us to easily generalize this to n dimensions, which we will pursue in the next class. For the time being, let us wind up by assuming that the Gaussian pdf in two dimensions is given by the formula in (7), though we showed it only for a particular method of generating the vector (X_1, X_2).

5 Covariance Matrices

We have encountered the matrix K while dealing with Gaussian random vectors. Clearly K is related to the moments of the participating random variables. What is the significance of K? Is it merely an end-product of the manipulations on joint probability density functions, as we saw in the last section? Turns out that the matrix K has some physical significance, and can be easily identified. It depends only on the individual and pair-wise relations of the participating random variables. The matrix K is popularly known as *covariance matrix*. Notice the connection to the word *variance* which we have introduced. It is kind of a *pair-wise* variance or inner product of random variables.

Definition 9 *For a random vector $\bar{X} = (X_1, \cdots, X_n)^T$ with $\mathbb{E}X_i^2 < \infty, 1 \le i \le n$, the covariance matrix K is defined as the outer product*

$$K = \mathbb{E}(\bar{X} - \bar{\mu})(\bar{X} - \bar{\mu})^T,$$

where $\bar{\mu} = \mathbb{E}[\bar{X}]$ is the mean vector.

Example 1 *Consider two independent Gaussian random variables X_1 and X_2 with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2), i = 1, 2$. Compute K of the vector $\bar{X} = (X_1, X_2)^T$.*

Solution: Notice that

$$\bar{X} = \bar{U} + \bar{\mu},$$

where $\bar{U} = (U_1, U_2)^T$ and $U_1 = X_1 - \mu_1$ and $U_2 = X_2 - \mu_2$. Thus $\mathbb{E}[\bar{U}] = 0$. Using this, we can write

$$\mathbb{E}[(\bar{X} - \bar{\mu})(\bar{X} - \bar{\mu})^T] = \mathbb{E}[\bar{U}\bar{U}^T].$$

Since $\mathbb{E}[U_1 U_2] = 0$,

$$K = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

■

In the last example, notice that the covariance matrix is diagonal. A diagonal matrix implies the lack of pairwise covariance. This is different, and in fact weaker, in several cases, from saying that X_1 and X_2 are independent.

Definition 10 *Random variables X_1 and X_2 are called **uncorrelated** if*

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2].$$

We have already showed that if X_1 and X_2 are independent, the $\mathbb{E}[g_1(X_1) g_2(X_2)] = \mathbb{E}[g_1(X_1)]\mathbb{E}[g_2(X_2)]$. Thus independence implies uncorrelatedness, and thus the former is stronger. The reverse is not always true, but for the important class of Gaussian Random variables independence and uncorrelatedness amount to the same notion. Before we prove such a statement, we need to first show that linear combinations of independent Gaussians will result in Gaussians. In other words, Gaussians form an invariant distribution under linear transformations. While this can be proved in many ways, perhaps the simplest proof uses an analog from the generating functions of discrete random variables, which we call characteristic functions in the real-valued case.

6 Characteristic Functions

Recall that we defined the $Z-$transform of the probability law of a discrete random variable as its generating function. The transform there is a polynomial representation. In continuous valued case, we cannot simply use polynomials and their coefficients, but the more general framework of Fourier (or two-sided Laplace) transform is required. We call this the characteristic function of random variable X.

Definition 11 *The characterestic function of a random variable X is defined as*

$$\mathbb{E}[e^{sX}],$$

where s takes values in the complex plane.

Observe that for the discrete case, we get back our generating function by substituting $z = e^s$. For the continuous valued cases which admit a density, it is sufficient to consider the Fourier transform of the pdf, and take $s = -j\omega$ to conclude our results. We have learnt in signals and systems that for a large class of functions, the Fourier transform completely specifies the function, for example integrable functions. Even otherwise, we know the formalism to determine the function almost everywhere from its Fourier Transform. Thus, we assume for the rest of the section that the characteristic function uniquely specifies a random variable. In other words, two random variables having the same characteristic function are considered to be identical.

To show the power of this transformations, let us find the probability distribution of the sum of two independent random variables.

Example 2 *Compute the pdf of $Y = X_1 + X_2$, where X_1 and X_2 are independent random variables admitting respective densities f_{X_1} and f_{X_2}.*

Solution: We will find the charactertistic function $\mathbb{E}[e^{sY}]$.

$$\begin{aligned}\mathbb{E}[e^{sY}] &= \mathbb{E}e^{s(X_1+X_2)}\\ &= \mathbb{E}e^{sX_1}\mathbb{E}e^{sX_2}.\end{aligned}$$

The last step used the independence assumptions of X_1 and X_2. We also know that multiplications in the frequency domain corresponds to convolutions in time domain. Thus, taking inverse Fourier Transforms, we get

$$f_Y = f_{X_1} \star f_{X_2},$$

a result which we already learnt. ■

7 Characteristic Functions and Gaussians

Let us now show that linear combination of independent Gaussians are always Gaussians. To this end, let us find the characteristic function of a Gaussian random variable.

Theorem 8 *Consider $X \sim \mathcal{N}(\mu, \sigma^2)$, then*

$$\mathbb{E}[e^{sX}] = e^{s\mu + \frac{1}{2}s^2\sigma^2}.$$

Proof: Integrate and apply the fact that the pdf of a random variable sums to one. ■

Theorem 9 *Let $\bar{X} = (X_1, \cdots, X_n)^T$ be a random vector containing independent Gaussian entries of mean $\mu_i, 1 \leq i \leq n$ and variance $\sigma_i^2, 1 \leq i \leq n$ respectively. Then for any $a \in \mathbb{R}^n$, the random variable $a^T \bar{X} = \sum_i a_i X_i$ is Gaussian distributed with mean $\sum a_i \mu_i$ and variance $\sum a_i^2 \sigma_1^2$.*

Proof: Let us find the generating function of $a^T \bar{X}$. Using the independence of X_i,

$$\begin{aligned}
\mathbb{E}[e^{sa^T \bar{X}}] &= \prod_{i=1}^{n} \mathbb{E} e^{s a_i X_i} \\
&= \prod_{i=1}^{n} e^{s a_i \mu + \frac{1}{2} s^2 a_i^2 \sigma^2} \\
&= e^{+s \sum_{i=1}^n a_i \mu_i + \frac{1}{2} s^2 \sum_{i=1}^n a_i^2 \sigma_i^2}.
\end{aligned}$$

This expression corresponds to a Gaussian random variable $\mathcal{N}(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2)$.

Theorem 10 *Consider a random vector $\bar{X} = (X_1, X_2)^T$ with uncorrelated entries. If $\bar{X}$ is jointly Gaussian with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2), i = 1, 2$ then X_1 and X_2 are independent.*

Proof: The covariance matrix of $\bar{X}$ is of the form

$$K = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2, \end{bmatrix}$$

since they are uncorrelated. Recall that the Gaussian joint pdf is completely determined by the mean vector $\bar{\mu}$ and K, Thus

$$\begin{aligned}
f_{X_1, X_2}(x_1, x_2) &= \frac{1}{\sqrt{\det(2\pi K)}} e^{-\frac{1}{2}(\bar{x}-\bar{\mu})^T K^{-1} (\bar{x}-\bar{\mu})} \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2} \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2} \\
&= f_X(x_1) f_{X_2}(x_2).
\end{aligned}$$

Comparing this with the joint pdf, it is evident that $P(X_1 \leq x_1, X_2 \leq x_2) = P(X_1 \leq x_1) P(X_2 \leq x_2)$ and thus X_1 and X_2 are independent. ■

Let us now precisely define what is meant by a Gaussian random vector, that we can extend our results to arbitrary dimensions.

Definition 12 *A random vector $(X_1, \cdots, X_n)$ is jointly Gaussian if $a^T \bar{X}$ is Gaussian for all non-trivial $a \in \mathcal{R}^n$.*

Example 3 *Consider two independent zero mean Gaussian random variables X_1 and X_2 with variances σ_1^2 and σ_2^2 respectively. Let*

$$\begin{bmatrix} U_1 & U_2 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

Show that (X_1, X_2) is not jointly Gaussian.

Solution: Observe that $2U_1 - U_2 = 0$, which is not a strict Gaussian random variable(it is a trivial random variable). ■

7.1 Multi-Dimensional Gaussian

Definition 13 *A jointly Gaussian vector $\bar{X} = (X_1, \cdots, X_n)^T$ is specified by the pdf*

$$f_{\bar{X}}(\bar{x}) = \frac{1}{\sqrt{\det(2\pi K)}} e^{-\frac{1}{2}(\bar{x}-\bar{\mu})^T K^{-1}(\bar{x}-\bar{\mu})},$$

where K is the covariance matrix and $\bar{\mu} = \mathbb{E}[\bar{X}]$. We will say $\bar{X} \sim \mathcal{N}(\bar{\mu}, K)$.

It is easy to see that both uncorrelated as well as independence implies the same notion for multi-dimensional jointly Gaussian random vectors, as the matrix K becomes diagonal in each case.

8 Transformation of Random Variables

Several practically important Random variable can be derived from Gaussians, by means of different functions and transformations. The characteristic function framework allows seamless conversions between various distributions obtained by transformation of random variables.

Theorem 11 *If the random variable X admits a pdf, and $Y = aX$, then*

$$f_Y(y) = \frac{1}{|a|} f_X(\frac{x}{a}).$$

Proof: Computing the characteristic function,

$$\begin{aligned}
\mathbb{E}e^{sY} &= \mathbb{E}e^{saY} \\
&= \int_{\mathbb{R}} e^{asx} f_X(x) dx \\
&= \int_{\mathbb{R}} e^{su} f_X(\frac{u}{a}) \frac{du}{|a|} \\
&= \int_{\mathbb{R}} e^{su} \frac{f_X(\frac{u}{a})}{|a|} du
\end{aligned}$$

By Inverse Fourier Transform, we can identify,

$$f_Y(y) = \frac{1}{|a|} f_X(\frac{x}{a}).$$

■

The important point to realize in the example is that the Jacobian matrix $J = \frac{dy}{dx} = a$. So the integral transformation takes the form

$$\mathbb{E}e^{sY} = \int_{\mathbb{R}} e^{su} \frac{f_X(\frac{u}{a})}{|\det(J)|} du.$$

This is the key in doing change of variables in integration. If there are multiple variables to be changed, we have to compute the Jacobian matrix J and divide by the absolute value of $\det(J)$. Let us do an example.

Example 4 *Show that if $Y = AX$,*

$$f_Y(y) = \frac{f_X(A^{-1}y)}{|\det A|}.$$

Solution: Comparing with the last theorem, the above result is akin to saying that the Jacobian matrix is A . Recall that $\det(A) = \det(A^T)$ for the $n \times n$ matrix A. We follow the same route as the scalar case, except that the Fourier transform also contains n parameters/frequency variables $s_1, s_2, \cdots s_n$. Thus the charactersitic function is $\mathbb{E}[e^{s^T Y}]$, where $s = (s_1, \cdots, s_n)^T$.

$$\mathbb{E}[e^{s^T Y}] = \int f_X(x) e^{s^T Ax} dx$$

Let $u = Ax$, then $\frac{du}{dx} = A^T$ and the Jacobian $J = A$. This is because $J_{ij} = \frac{\delta u_i}{\delta x_j}$ by definition of the Jacobian, while the matrix derivative of Ax is conventionally defined as A^T. In any case, $\det J = \det(J^T)$ (i.e. transpose or not will not make a difference to the computation). Thus,

$$\mathbb{E}[e^{s^T Y}] = \int \frac{f_X(A^{-1}u)}{|\det(A)|} e^{s^T u} du.$$

Clearly, the characteristic function appears as the Fourier transform of $\frac{f_X(A^{-1}u)}{|\det(A)|}$, validating our claim.

■

Theorem 12 *If* $Y = X^2$, *then*

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left(f_X(\sqrt{y}) + f_X(-\sqrt{y}) \right).$$

Proof: We have shown the direct computation in class. We can also use the generating function framework.

9 Markov's Inequality

Recall the Markov's inequality for the discrete random variables. An exact analog holds for continuous valued random variables too. We will state a more general version.

Theorem 13 *For a non-negative random variable* X,

$$P(X > a) \le \frac{\mathbb{E}[X]}{a}, \ a > 0.$$

Proof: The proof follows exactly as in the discrete case, in particular

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X \mathbb{1}_{\{X \le a\}} + X \mathbb{1}_{\{X > a\}}] \\ &\ge \mathbb{E}[X \mathbb{1}_{\{X > a\}}] \\ &\ge \mathbb{E}[a \mathbb{1}_{\{X > a\}}] \\ &= a\mathbb{E}[\mathbb{1}_{\{X > a\}}] = aP(X > a).\end{aligned}$$

Many other theorems and inequalities related to the expectation carry over from the discrete case to the continuous valued ones. We may not repeat each of them, the reader can check the proofs in the discrete case, to find out whether it can be generalized.

10 Conditional Probability

As the name implies, there are at least two random variables involved, let us denote it as X and Y. Let us start with a familiar idea of probabilities of the joint events of the form i.e. $\{X \le x\}$ and $\{Y \le y\}$. We will use this later to define conditional probabilities. We know that

$$P(X \le x | Y \le y) = \frac{P(X \le x, Y \le y)}{P(Y \le y)} \mathbb{1}_{\{P(Y \le y) \ne 0\}}.$$

In the all-discrete case, we used $P(Y = y)$ instead of $P(Y \le y)$ in the above expression, to define the conditional cumulative distribution. However, this cannot be done if Y is absolutely continuous, as $P(Y = y) = 0$ uniformly for all y. So we will define four separate versions of conditional probability depending on whether X and Y are continuous or not. Though we say different versions, all of them have the same theme, with appropriate replacement of distributions by densities. The first case is when X and Y are discrete, which is already familiar.

1. $\underline{X\text{– discrete}, Y\text{– discrete}}$

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \mathbb{1}_{\{P(Y = y) > 0\}}.$$

2. $\underline{X\text{– continuous}, Y\text{– discrete}}$

   Here we will replace the event $\{X = x\}$ by $\{X \le x\}$

$$P(X \le x | Y = y) = \int_{-\infty}^{x} f_{X|Y}(u|y) du \mathbb{1}_{\{P(Y = y) > 0\}},$$

   where we used the notation $f_{X|Y}(x|y)$ to denote the density function of X given $Y = y$. Notice that

$$\begin{aligned} P(X \le x | Y = y) &= \frac{P(X \le x, Y = y)}{P(Y = y)} \mathbb{1}_{\{P(Y = y) > 0\}} \\ &= \mathbb{1}_{\{P(Y = y) > 0\}} \frac{1}{P(Y = y)} \sum_{v = y} \int_{-\infty}^{x} f_{XY}(u, v) du \\ &= \int_{-\infty}^{x} \frac{1}{P(Y = y)} f_{XY}(u, y) \mathbb{1}_{\{P(Y = y) > 0\}} du \end{aligned}$$

   In this case, it makes sense to write

$$f_{X|Y}(u|y) = \frac{f_{XY}(u, y)}{P(Y = y)} \mathbb{1}_{\{P(Y = y) > 0\}}. \tag{9}$$

   Since the above case occurs frequently in the examples of electrical engineering, we will conveniently denote $f_{X|Y}(u|k)$ as $f_k(x)$, i.e. the conditional density of X given $Y = k$. To check the consistency of our definition, observe that

$$\begin{aligned} \int_{\mathbb{R}} f_{X|Y}(u|y) du &= \int \frac{f_{XY}(u, y)}{P(Y = y)} \mathbb{1}_{\{P(Y = y) > 0\}} du \\ &= \frac{\mathbb{1}_{\{P(Y = y) > 0\}}}{P(Y = y)} \int f_{XY}(u, y) du \\ &= \frac{\mathbb{1}_{\{P(Y = y) > 0\}}}{P(Y = y)} P(Y = y) \\ &= \mathbb{1}_{\{P(Y = y) > 0\}}. \end{aligned}$$

Thus, our definition gives a valid probability measure for all $y : P(Y = y) > 0$. Also, the marginal distribution of X becomes

$$P(X \leq x) = \sum_k P(Y = k) \int_{-\infty}^{x} f_k(x)dx.$$

3. <u>$X-$ discrete, $Y-$continuous</u>

   We will reverse engineer a consistent definition of conditional distribution from the last two cases. Specifically, let us define

   $$P(X = i|Y = y) = \frac{f_{XY}(i,y)}{f_Y(y)} \mathbb{1}_{\{f_Y(y)>0\}}. \tag{10}$$

   Clearly

   $$\begin{aligned}\sum_i P(X = i|Y = y) &= \sum_i \frac{f_{XY}(i,y)}{f_Y(y)} \mathbb{1}_{\{f_Y(y)>0\}} \\ &= \frac{f_Y(y)}{f_Y(y)} \mathbb{1}_{\{f_Y(y)>0\}} \\ &= \mathbb{1}_{\{f_Y(y)>0\}}.\end{aligned}$$

   Thus, our definition gives a valid probability measure for all $y : f_Y(y) > 0$. The conditional probability that we defined also takes the convenient form,

   $$.\frac{f_{XY}(i,y)}{f_Y(y)} = \frac{P(X = i)f_i(y)}{f_Y(y)} = \frac{P(X = i)f_i(y)}{\sum_i P(X = i)f_i(y)}, \tag{11}$$

   whenever the quantities are well-defined. Recall the definition of the conditional density $f_i(y)$ from the previous case.

4. <u>$X-$ continuous, $Y-$ continuous</u>

   Let us define

   $$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \mathbb{1}_{\{f_Y(y)>0\}}. \tag{12}$$

   It is simple to check that

   $$\int_{\mathbb{R}} f_{X|Y}(x|y)dx = \mathbb{1}_{\{f_Y(y)>0\}}.$$

   Thus, it is a valid density function.

10.1 Conditional Distribution of Random Vectors

The definitions that we made are equally applicable to random vectors, which are of the mixed type, i.e. they may have some elements discrete and rest continuous. In each case, we can modify the appropriate definition and get consistent results.

11 Conditional Expectation

Expectation with respect to a conditional distribution is known as conditional expectation. Since we defined the conditional distribution for 4 separate cases, the conditional expectation has to be evaluated accordingly in these cases. For generality, we will denote the conditional distribution that we introduced in the last section by $\Pi(x|y)$. We will mention the general framework here.

Definition 14 *Consider a function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, which is either non-negative or integrable. The function*

$$\Psi(y) = \mathbb{E}[g(X,Y)|Y = y]$$

is known as the conditional expectation of $g(X,Y)$ given $Y = y$, where the expectation is evaluated with respect to $\Pi(x|y)$.

Observe that this definition can be easily specialized to each of the cases that we dealt. The most important case in our discussion is the last one, where it reads,

$$\Psi(y) = \int_{\mathbb{R}} g(x,y)\Pi(x|y)dx$$

in which $\Pi(x|y)$ is same as the function $f_{X|Y}(x|y)$ given in (12). On the other hand, when X is discrete and Y continuous, we can write

$$\Psi(y) = \sum_i g(i,y)\Pi(i|y),$$

where $\Pi(i|y)$ is taken as per (10), or the more convenient form in (11).

Example 5 *Let (X_1, X_2) be a zero mean jointly Gaussian random vector with covariance matrix*

$$K = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Find $\mathbb{E}[X_1|X_2]$ and $\mathbb{E}[X_1^2|X_2]$.

Solution: Notice that we need to specialize our definition of conditional distribution. In particular, since both RVs are continuous, let us look at their joint density.

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{\sqrt{\det(2\pi K)}} e^{-\frac{1}{2}x^T K^{-1} x},$$

where $x = (x_1, x_2)^T$. The marginal distribution of X_2 is a Gaussian (since the given vector is jointly Gaussian), which we can identify as

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{x_2^2}{2\sigma_2^2}}.$$

The conditional density becomes

$$\frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} = \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_1^2}} e^{-\frac{1}{2(1-\rho^2)\sigma_1^2}(x_1 - \rho\frac{\sigma_1}{\sigma_2}x_2)^2}.$$

Observe that the conditional density is nothing but a Gaussian density function with mean $\rho\frac{\sigma_1}{\sigma_2}x_2$ and variance $(1-\rho^2)\sigma_1^2$. Thus, it is easy to identify

$$\mathbb{E}[X_1|X_2] = \rho\frac{\sigma_1}{\sigma_2}x_2 \text{ and } \mathbb{E}[X_1^2|X_2] = (1-\rho^2)\sigma_1^2 + \rho^2\frac{\sigma_1^2}{\sigma_2^2}x_2^2.$$

■

12 Other results

It is clear that $\Psi(Y)$ defined above is a random variable taking values in $\bar{\mathbb{R}}$ whenever $\mathbb{E}|g(X,Y)| < \infty$, and then it is meaningful to talk about its expectation. We have performed such computations for the discrete case. Many results from the conditional expectation for discrete cases have exact analogs in the general case. Here we list a few of them, whose proofs are straightforward, whenever the quantities involved are finite or non-negative. For example

1. $$\mathbb{E}\left[\mathbb{E}[g(X,Y)|Y = y]\right] = \mathbb{E}g(X,Y) \tag{13}$$

2. $$\mathbb{E}[g(X)|X] = g(X) \tag{14}$$

3. If $X \perp\!\!\!\perp Y$,
$$\mathbb{E}_y\mathbb{E}_x[g(X)|Y] = \mathbb{E}g(X).$$

4. Wald's Inequality.

Exercise 3 *Using the definition of conditional probabilities, prove each of these expressions.*