EE 520: Random Processes **Fall 2021**

Lecture 5
Continuous Random Variables and Detection Theory Primer

Instructor Name: John Lipor

Recommended Reading: Pishro-Nik: 4.1 (skip 4.1.3), 4.2, 6.1.3, 8.4, 9.1.8; Gubner: 4.1 - 4.3

1 Densities and Probabilities

We now consider random variables that can take values in an uncountable set, which we call *continuous* RVs.

Definition 1. A random variable is called **continuous** if

$$P(X \in B) = \int_B f_X(t)dt = \int_{-\infty}^{\infty} \mathbb{1}_B(t) f_X(t)dt,$$

where $\mathbb{1}_B(x)$ is the indicator function taking the value 1 if $x \in B$ and 0 otherwise and $f_X(t)$ is called the **probability density function** (PDF).

We're often interested in sets B that have the form $[a, b]$, so we typically consider tasks such as computing

$$P(a \leq X \leq b) = \int_a^b f_X(t)dt.$$

Note: The numerical value $f_X(t)$ should **not** be viewed as a probability. For continuous RVs, it holds that

$$P(X = t) = 0 \quad \forall t \in \mathbb{R},$$

i.e., the probability that a continuous RV takes on any *single* value is zero. Instead, you should think of $f_X(t)dt$ as the element of probability

$$P(t \leq X \leq t + dt).$$

1.1 Common Densities

- **uniform**: Outcome "random" but lies in some known interval. This has the PDF

$$f_X(t) = \begin{cases} \frac{1}{b-a} & a \leq t \leq b \\ 0 & \text{otherwise} \end{cases}$$

- **exponential**: Non-negative RVs that decay over time, having PMF

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- **Laplace**: Double-sided exponential with PMF

$$f_X(t) = \frac{\lambda}{2} e^{-\lambda |t|}$$

- **Gaussian**: Also known as the "normal" distribution and is probably the most common density. Has PDF
$$f_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right),$$
where μ is the mean and $\sigma^2 > 0$ is the variance. To indicate that X is a Gaussian RV, we write $X \sim \mathcal{N}(\mu, \sigma^2)$.

We will see in a few weeks that if we have RVs $X_1, X_2, \ldots$ all independent and standardized, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \to \mathcal{N}(0,1) \text{ as } n \to \infty$$

This is known as the central limit theorem. Continuous RVs are the main reason we need to deal with the σ-algebra, as illustrated by the following fact.

Fact 1. Let $\Omega = [0,1]$. No function P exists from the power set 2^Ω to the reals such that

1. $P([a,b]) = b - a, \quad \forall 0 \le a \le b \le 1$
2. the axioms of probability are satisfied.

The above fact states that if we insisted on measuring everything (i.e., if we ignored the need for σ-algebras), we would not be able to have the above common-sense definition of the uniform distribution. We may prove this fact later in the course if time allows.

Definition 2. If the joint density $f_{XY}(x,y)$ exists, we say X and Y are independent if

$$f_{XY}(x,y) = f_X(x) f_Y(y).$$

2 Expectation

For continuous RVs, expectation is analogous to the discrete case

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

for any function $g : \mathbb{R} \to \mathbb{R}$. In particular,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx.$$

The definitions of variance, correlation, and covariance are also analogous to the discrete case.

3 Transform Methods

A few transformations make computing functions of RVs easier sometimes. We will only use one of these in this course.

Definition 3. The **moment generating function** (MGF) of a RV X is defined as

$$M_X(s) = \mathbb{E}[e^{sX}] = \int_{-\infty}^{\infty} e^{sx} f_X(x) dx.$$

As the name indicates, the MGF can be used to compute moments, i.e., $\mathbb{E}[X^k]$ of a RV X. It holds that

$$\mathbb{E}[X^k] = \frac{\partial^k}{\partial s^k} M_X(s)|_{s=0}.$$

In words, the kth moment of X can be found by taking the kth derivative of the MGF and evaluating it at $s = 0$.

Example 1. Let $X \sim \mathcal{N}(0, 1)$. Then

$$\begin{aligned} M_X(s) &= \mathbb{E}[e^{sX}] \\ &= \int_{-\infty}^{\infty} e^{sx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= e^{s^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-s)^2/2} dx \\ &= e^{s^2/2}, \end{aligned}$$

where we have used the fact that the integrated term in the third line is the PDF of a $\mathcal{N}(s^2, 0)$ RV and therefore integrates to one. Now note that

$$\begin{aligned} \frac{\partial}{\partial s} M_X(s) = s e^{s^2/2} \quad &\implies \quad \mathbb{E}[X] = 0 \\ \frac{\partial^2}{\partial s^2} M_X(s) = s^2 e^{s^2/2} + e^{s^2/2} \quad &\implies \quad \mathbb{E}[X^2] = 1, \end{aligned}$$

both of which make sense given that we said X has zero mean and unit variance.

The MGF shows up frequently when bounding RVs, namely in Chernoff's bounding method, which we'll encounter in a few weeks. Bounding RVs becomes important in machine learning theory, when we want to bound the difference between the true and estimated error of an algorithm.

The MGF is also useful for handling sums of independent RVs.

Proposition 1. If X and Y are independent RVs and $Z = X + Y$, then

$$M_Z(x) = M_X(x) M_Y(s).$$

Proof. We apply the definition of the MGF and use independence.

$$\begin{aligned} M_Z(s) &= \int_{-\infty}^{\infty} e^{sz} f_Z(z) dz \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{s(x+y)} f_{XY}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{sx} e^{sy} f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} e^{sx} f_X(x) dx \int_{-\infty}^{\infty} e^{sy} f_Y(y) dy \\ &= M_X(s) M_Y(s). \end{aligned}$$

□

4 Detection Theory Primer

In a variety of electrical engineering topics ranging from communications to machine learning, we are interested in classifying signals/measurements in a statistically-optimal way.

Example 2. Transmit a signal X through a binary channel and receive signal Y. Given the received signal, what should we say was transmitted?

Example 3. Feature-label pairs (X, Y) are generated according to some distribution P_{XY}. Given a new feature vector $\hat{X}$, how should we decide the class label $\hat{Y}$?

We can use probability theory to optimize how we make these decisions. For now, we assume $X \in \{0, 1\}$, corresponding to the case of binary detection/classification.

4.1 Setup

Define the following terminology.

- H_0: null hypothesis ($X = 0$)
- H_1: alternate hypothesis ($X = 1$)
- D_0: decide H_0
- D_1: decide H_1

Errors occur when we either

1. are in the setting H_0 but decide $X = 1$ (false alarm)
2. are in the setting H_1 but decide $X = 0$ (missed detection),

so our total error is

$$E = (H_0 \cap D_1) \cup (H_1 \cap D_0).$$

Our goal is to design a detector to minimize the probability of error

$$\begin{aligned} P(E) &= P\left((H_0 \cap D_1) \cup (H_1 \cap D_0)\right) \\ &= P(H_0 \cap D_1) + P(H_1 \cap D_0) \\ &= P(D_1 \mid H_0)P(H_0) + P(D_0 \mid H_1)P(H_1). \end{aligned}$$

We will make our decision by separating the space into two regions, R_0 where we decide D_0 and R_1 where we decide D_1 (see Fig. 8.10 in Pishro-Nik). Minimizing the probability of error then amounts to minimizing the probability mass from each H_0 and H_1 that lies on the "wrong" side of the boundary. We want to minimize

$$P(E) = \int_{R_0} P(H_1) f(x \mid H_1) dx + \int_{R_1} P(H_0) f(x \mid H_0) dx.$$

Note that

$$1 = \int_{-\infty}^{\infty} f(x \mid H_1) dx = \int_{R_0} f(x \mid H_1) dx + \int_{R_1} f(x \mid H_1) dx,$$

so we can rewrite the probability of error as

$$\begin{aligned} P(E) &= \int_{R_1} P(H_0) f(x \mid H_0) dx + P(H_1) \int_{R_0} f(x \mid H_1) dx \\ P(E) &= \int_{R_1} P(H_0) f(x \mid H_0) dx + P(H_1) \left(1 - \int_{R_1} f(x \mid H_1) dx\right) \\ &= \int_{R_1} P(H_0) f(x \mid H_0) - P(H_1) f(x \mid H_1) dx + P(H_1). \end{aligned}$$

When we're optimizing anything in life, we should focus on the things we can change, as opposed to those we cannot (stoic philosophy quote of the lecture). Since $P(H_1)$ is fixed, we choose R_1 to minimize

$$\int_{R_1} P(H_0)f(x \mid H_0) - P(H_1)f(x \mid H_1)dx,$$

which is done by making R_1 the region over which

$$P(H_1)f(x \mid H_1) > P(H_0)f(x \mid H_0).$$

This leads to the **maximum a posteriori** (MAP or Bayesian) detection rule

$$P(H_0)f(x \mid H_0) \underset{}{\overset{H_0}{\gtrless}} P(H_1)f(x \mid H_1),$$

which indicates we choose H_0 if the top inequality holds and choose H_1 otherwise. We can also divide both sides by $f_X(x)$, which shows the relationship to the posterior probability more clearly and yields a second form of the MAP rule

$$f(H_0 \mid x) \underset{}{\overset{H_0}{\gtrless}} f(H_1 \mid x).$$

If we do not have any prior beliefs about the events H_0 or H_1, we can assume they are equal, which yields the **maximum likelihood** (ML) detection rule

$$f(x \mid H_0) \underset{}{\overset{H_0}{\gtrless}} f(x \mid H_1).$$

As a reminder to help you with the terminology, note that

- $P(H_0)$ is the **prior** probability that H_0 happens before any data is received
- $P(H_0 \mid x)$ is the **posterior** probability that H_0 happens now that we have the data x
- $P(x \mid H_0)$ is the **likelihood** that x takes some value given that H_0 occurred
- the MAP rule can be viewed through the **likelihood ratio** test, which under equal priors on H_0 and H_1 makes the MAP and ML detectors equivalent

$$\frac{f(x \mid H_1)}{f(x \mid H_0)} \underset{}{\overset{H_1}{\gtrless}} \frac{P(H_0)}{P(H_1)}.$$