Lisa Yan and Jerry Cain
CS109
Lecture Notes #17
October 21, 2020

Continuous Joint Distributions, Part II

Based on a chapter by Chris Piech and Lisa Yan

Pre-recorded lecture: Sections 1 and 2. **In-lecture**: Sections 3 and 4.
Not covered: Section 5.

1 Convolution: Sum of independent random variables

Remember how deriving the sum of two independent Poisson random variables was tricky? When we move into integral land, the concept of convolution still carries over, and once you get a handle on notation, then computing the sum of two independent, jointly continuous random variables becomes fun. For some definition of fun...

1.1 Independent Normals

Let's start with one common case that has a nice form but a difficult derivation:

For any two normal random variables $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ the sum of those two random variables is another normal: $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

The optional proof if you're totally excited about it is located in Section 5. It involves exponents, integrals, and completing the square (algebra throwback!).

1.2 General Independent Case

For two general independent random variables, you can calculate the CDF or the PDF of the sum of two random variables using the following convolution formulas:

$$F_{X+Y}(a) = P(X + Y \le a) = \int_{y=-\infty}^{\infty} F_X(a - y) f_Y(y) dy$$
$$f_{X+Y}(a) = \int_{y=-\infty}^{\infty} f_X(a - y) f_Y(y) dy$$

These is a direct analogy to the discrete case where you replace the integrals with sums and change notation for CDF and PDF.

1.3 Example: Sum of Independent Uniforms

What is the PDF of $X + Y$ for independent uniform random variables $X \sim \text{Uni}(0,1)$ and $Y \sim \text{Uni}(0,1)$? First plug in the equation for general convolution of independent random variables:

$$f_{X+Y}(a) = \int_{y=0}^{1} f_X(a-y) f_Y(y) dy$$

$$f_{X+Y}(a) = \int_{y=0}^{1} f_X(a-y) dy \qquad \text{because } f_Y(y) = 1$$

It turns out that is not the easiest thing to integrate. By trying a few different values of a in the range $[0,2]$ we can observe that the PDF we are trying to calculate is discontinuous at the point $a = 1$ and thus will be easier to think about as two cases: $a < 1$ and $a > 1$. If we calculate f_{X+Y} for both cases and correctly constrain the bounds of the integral we get simple closed forms for each case:

$$f_{X+Y}(a) = \begin{cases} a & \text{if } 0 < a \leq 1 \\ 2 - a & \text{if } 1 < a \leq 2 \\ 0 & \text{else} \end{cases}$$

2 Conditional Distributions (Continuous case)

The conditional probability density function might look a bit wonky, but it works!

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

At first glance, the conditional density function seems to violate any notion of stoichiometric units of probability. Let us verify this with our understanding of discrete probability. Recall that for tiny epsilon ε, we can approximate $P(|X - x| \leq \frac{\varepsilon}{2}) = P(x - \frac{\varepsilon}{2} \leq X \leq x + \frac{\varepsilon}{2}) = \int_{x-\frac{\varepsilon}{2}}^{x+\frac{\varepsilon}{2}} f_X(a) da \approx f_X(x)\varepsilon$. This extends to the joint variable case: $P(|X - x| \leq \frac{\varepsilon_X}{2}, |Y - y| \leq \frac{\varepsilon_y}{2}) \approx f_{X,Y}(x,y)\varepsilon_X \varepsilon_Y$.

$$P\left(|X - x| \leq \frac{\varepsilon_X}{2} \middle| |Y - y| \leq \frac{\varepsilon_Y}{2}\right) = \frac{P(|X - x| \leq \frac{\varepsilon_X}{2}, |Y - y| \leq \frac{\varepsilon_Y}{2})}{P(|Y - y| \leq \frac{\varepsilon_Y}{2})} \qquad \text{(def. cond. prob.)}$$

$$\approx \frac{f_{X,Y}(x,y)\varepsilon_X \varepsilon_Y}{f_Y(y)\varepsilon_Y} = \frac{f_{X,Y}(x,y)}{f_Y(y)} \varepsilon_X = f_{X|Y}(x|y)\varepsilon_X$$

If you're interested, the conditional cumulative distribution function for the continuous case is

$$F_{X|Y}(a|y) = P(X \leq a | Y = y) = \int_{-\infty}^{a} f_{X|Y}(x|y) dx$$

Furthermore, conditional expectation in the continuous case is a direct analogy to what we saw in the discrete case. Let X and Y be jointly continuous random variables. We define the conditional expectation of X given $Y = y$ to be:

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx$$

3 Exercises

3.1 Basketball games

Two teams have ELO scores $A_W \sim \mathcal{N}(1657, 200^2)$ and $A_B \sim \mathcal{N}(1470, 200^2)$ that represent their respectively ability to play well on any given day. What is $P(A_W > A_B)$, the probability that the first team's ability is greater than the second team's ability?

Solution: We note that we can transform our event into $P(A_W - A_B > 0$. Then $A_W - A_B$ is a sum of two independent Normal random variables, A_W and $-A_B \sim \mathcal{N}(-1470, (-1)^2 200^2 = 200^2)$. Therefore $A_W - A_B \sim \mathcal{N}(1657 - 1470 = 187, 200^2 + 200^2 \approx 283^2)$. We then compute $P(A_W - A_B > 0) \approx 1 - \Phi(\frac{0-187}{283}) \approx 0.7454$.

3.2 Example: Linear Transform of Normal versus Sum of Independent Normals

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = X + X$. Since $Y = 2X$, by linearity, $Y \sim \mathcal{N}(2\mu, 4\sigma^2)$.

It may be tempting to think that Y is a sum of independent Normal RVs to claim that Y has a variance of $\sigma^2 + \sigma^2 = 2\sigma^2$, but note that X is not independent of X! So Y is Normal not because sums of independent Normal RVs are Normal—but rather because a linear transform of a Normal RV is Normal.

3.3 Example: Virus Infections

Suppose that you are working with a health organization to plan a response to the initial conditions of a virus. There are two exposed groups: The first group has 200 people, where each person is infected independently with the virus with probability $p_1 = 0.1$. The second group has 100 people, where each person is infected independently with probability $p_2 = 0.4$. What is the *approximate* probability that the total number of people infected is more than 55 people?

Solution: Let A and B be random variables representing the number of people infected in the first and second groups, respectively. We would like to compute $P(A+B \geq 55)$, where $A \sim \text{Bin}(200, 0.1)$ and $B \sim \text{Bin}(100, 0.4)$. We note that while $A + B$ is sum of independent Binomial RVs, we will need to use convolution to calculate its distribution, since A and B have different probabilities of success per trial.

Instead, we approximate A and B as Normal random variables X and Y, respectively, and subsequently $P(A + B \geq 55) \approx P(X + Y \geq 54.5)$ (with continuity correction). $A \approx X \sim \mathcal{N}(20, 18)$ and $B \approx Y \sim \mathcal{N}(40, 24)$, where a Normal approximation is appropriate because $\text{Var}(A) = 200(0.1)(0.9) > 10$ and $\text{Var}(B) = 100(0.4)(0.6) > 10$. Then $X+Y$ is a sum of independent Normal random variables; therefore $X + Y \sim \mathcal{N}(60, 42)$, and $P(X + Y \geq 54.5) = 1 - \Phi(\frac{54.5-60}{\sqrt{42}}) \approx 0.8023$.

3.4 Example: Conditional densities

Let X and Y be continuous random variables with joint PDF $f_{X,Y}(x, y) = \frac{12}{5}x(2 - x - y)$ in the support $0 < x, y < 1$ (and zero otherwise). Are X and Y independent?

Solution: We can compute the conditional density of X given $Y = y$, where $0 < x, y < 1$ as

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx} = \frac{\frac{12}{5}x(2-x-y)}{\int_0^1 \frac{12}{5}x(2-x-y)dx}$$
$$= \frac{x(2-x-y)}{\left[x^2 - x^3/3 - x^2y/2\right]_0^1} = \frac{x(2-x-y)}{2/3 - y/2} = \frac{6x(2-x-y)}{4-3y}$$

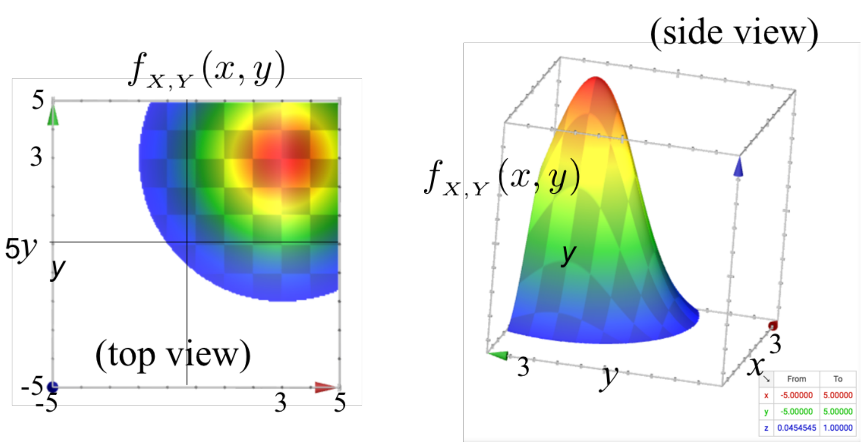
4 Example: 2-D Tracking

In this example we are going to explore the problem of tracking an object in 2D space. The object exists at some (x, y) location, however we are not sure exactly where! Thus we are going to use random variables X and Y to represent location.

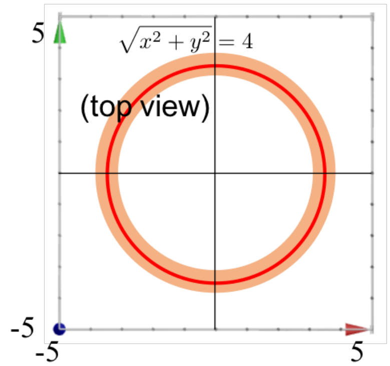
We have a prior belief about where the object is. In this example our prior both X and Y as normals which are independently distributed with mean 3 and variance 4. First let's write the prior belief as a joint probability density function

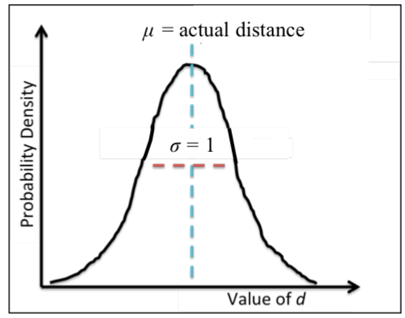
$$f(X=x, Y=y) = f(X=x) \cdot f(Y=y) \qquad \text{In the prior X and Y are independent}$$
$$= \frac{1}{\sqrt{2\cdot 4\cdot \pi}} \cdot e^{-\frac{(x-3)^2}{2\cdot 4}} \cdot \frac{1}{\sqrt{2\cdot 4\cdot \pi}} \cdot e^{-\frac{(y-3)^2}{2\cdot 4}} \qquad \text{Using the PDF equation for normals}$$
$$= K_1 \cdot e^{-\frac{(x-3)^2+(y-3)^2}{8}} \qquad \text{All constants are put into } K_1$$

This combinations of normals is called a bivariate distribution. Here is a visualization of the PDF of our prior.



The interesting part about tracking an object is the process of updating your belief about it's location based on an observation. Let's say that we get an instrument reading from a sonar that is sitting on the origin. The instrument reports that the object is 4 units away. Our instrument is not perfect: if the true distance was t units away, than the instrument will give a reading which is normally distributed with mean t and variance 1. Let's visualize the observation:





Based on this information about the noisiness of our prior, we can compute the conditional probability of seeing a particular distance reading D, given the true location of the object X, Y. If we knew the object was at location (x, y), we could calculate the true distance to the origin $\sqrt{x^2+y^2}$ which would give us the mean for the instrument Gaussian:

$$f(D = d|X = x, Y = y) = \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \cdot e^{-\frac{\left(d-\sqrt{x^2+y^2}\right)^2}{2\cdot 1}} \quad \text{Normal PDF where } \mu = \sqrt{x^2+y^2}$$

$$= K_2 \cdot e^{-\frac{\left(d-\sqrt{x^2+y^2}\right)^2}{2\cdot 1}} \quad \text{All constants are put into } K_2$$

How about we try this out on actual numbers. How much more likely is an instrument reading of 1 compared to 2, given that the location of the object is at (1, 1)?

$$\frac{f(D = 1|X = 1, Y = 1)}{f(D = 2|X = 1, Y = 1)} = \frac{K_2 \cdot e^{-\frac{\left(1-\sqrt{1^2+1^2}\right)^2}{2\cdot 1}}}{K_2 \cdot e^{-\frac{\left(2-\sqrt{1^2+1^2}\right)^2}{2\cdot 1}}} \quad \text{Substituting into the conditional PDF of D}$$

$$= \frac{e^0}{e^{-1/2}} \approx 1.65 \quad \text{Notice how the } K_2 \text{ cancel out}$$

At this point we have a prior belief and we have an observation. We would like to compute an updated belief, given that observation. This is a classic Bayes' formula scenario. We are using joint continuous variables, but that doesn't change the math much, it just means we will be dealing with densities instead of probabilities:

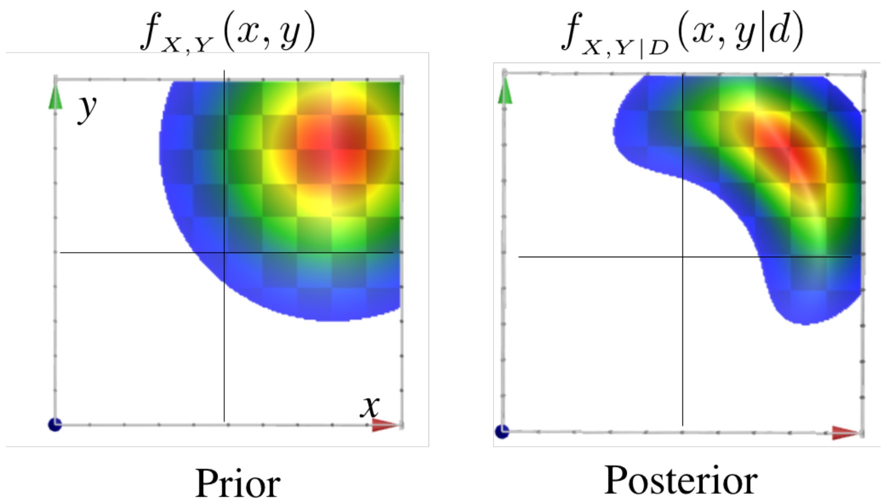
$$f(X = x, Y = y|D = 4) = \frac{f(D = 4|X = x, Y = y) \cdot f(X = x, Y = y)}{f(D = 4)} \quad \text{Bayes using densities}$$

$$= \frac{K_1 \cdot e^{-\frac{[4-\sqrt{x^2+y^2})^2]}{2}} \cdot K_2 \cdot e^{-\frac{[(x-3)^2+(y-3)^2]}{8}}}{f(D = 4)} \quad \text{Substituting for prior and update}$$

$$= \frac{K_1 \cdot K_2}{f(D = 4)} \cdot e^{-\left[\frac{[4-\sqrt{x^2+y^2})^2]}{2} + \frac{[(x-3)^2+(y-3)^2]}{8}\right]} \quad f(D = 4) \text{ is a constant w.r.t. } (x, y)$$

$$= K_3 \cdot e^{-\left[\frac{(4-\sqrt{x^2+y^2})^2}{2} + \frac{[(x-3)^2+(y-3)^2]}{8}\right]} \quad K_3 \text{ is a new constant}$$

Wow! That looks like a pretty interesting function! You have successfully computed the updated belief. Let's see what it looks like. Here is a figure with our prior on the left and the posterior on the right: How beautiful is that! Its like a 2D normal distribution merged with a circle. But wait, what



about that constant! We do not know the value of K_3 and that is not a problem for two reasons: the first reason is that if we ever want to calculate a relative probability of two locations, K_3 will cancel out. The second reason is that if we really wanted to know what K_3 was, we could solve for it.

This math is used every day in millions of applications. If there are multiple observations the equations can get truly complex (even worse than this one). To represent these complex functions often use an algorithm called particle filtering.

5 Proof of Sum of Independent Normals

There are also several interesting proofs for this property—that the sum of independent normal random variables is also normally distributed—which either involve geometry, Fourier transforms, or characteristic functions, but in this section we borrow largely from the Ross textbook (specifically, section 6.3.3 in the 10th edition). The proof is performed in two parts.

We first prove that two independent standard normal random variables $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 1)$ will have a sum $X + Y \sim \mathcal{N}(0, 2)$. We focus on evaluating the expression inside the integral when convolving two independent continuous random variables:

$$f_{X+Y}(a) = \int_{y=-\infty}^{\infty} f_X(a-y) f_Y(y)\, dy,$$

where

$$f_X(a-y) f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-(a-y-0)^2/2} \frac{1}{\sqrt{2\pi}} e^{-(y-0)^2/2} = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2\pi}} e^{-((a-y)^2+y^2)/2}.$$

Next, we split the exponential component into two squares—one involving y and one not—by completing the square (yes, it's happening!):

$$\begin{aligned}\frac{-1}{2}\left[(a-y)^2+y^2)\right] &= \frac{-1}{2}\left[a^2-2ay+2y^2\right] = -1\left[y^2-ay+\frac{a^2}{2}\right]\\ &= -1\left[\left(y-\frac{a}{2}\right)^2+\frac{a^2}{2}-\frac{a^2}{4}\right] = \frac{-1}{2}\left[2\left(y-\frac{a}{2}\right)^2+\frac{a^2}{2}\right].\end{aligned}$$

We return back to our expression for convolving X and Y:

$$\begin{aligned}f_{X+Y}(a) &= \int_{y=-\infty}^{\infty} f_X(a-y)f_Y(y)\,dy\\ &= \int_{y=-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{2\pi}}e^{(-1/2)\left[2(y-a/2)^2+a^2/2\right]}\,dy\\ &= \frac{1}{\sqrt{2\pi}}e^{-(1/2)\left[a^2/2\right]}\int_{y=-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-(1/2)\left[(y-a/2)^2/(1/2)\right]}\,dy\\ &= \frac{1}{\sqrt{2\pi}\sqrt{2}}e^{-(1/2)\left[a^2/2\right]}\int_{y=-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sqrt{1/2}}e^{-(1/2)\left[(y-a/2)^2/(1/2)\right]}\,dy\end{aligned}$$

The expression in the integral is a normal PDF on y given a (specifically, $\mathcal{N}(a/2, 1/2)$) and therefore the PDF will integrate to 1, and we are left with

$$f_{X+Y}(a) = \frac{1}{\sqrt{2\pi}\sqrt{2}}e^{-(1/2)\left[a^2/2\right]},$$

and therefore $X+Y \sim \mathcal{N}(0,2)$.

Next, we prove the general case, where $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2)$ have sum $X+Y \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$. Note that all linear transforms of normal random variables are normal, and therefore $X+Y$ can be shown to be the linear transform of the sum of two standard normal random variables, whose sum is distributed as $\mathcal{N}(0,2)$. We leave the details of the second part to you. Have fun!

P.S. – Wondering how to derive the linearity property of normal random variables? Check out Section 5.4 of the Ross textbook for the proof.