The Jacobian Formula: functions are linear if you look really close

Notational remark: The bolded variables are either matrices or vectors; I like to do that to visually remind myself what is what exactly. This will be a little confusing because usually bolded uppercase letters are matrices, lower case are vectors, but here I'm also adding *random vectors* as bolded upper-case letters. Also, $|\cdot|$, when applied to a matrix, is the *absolute value of the determinant.*

The multivariate derived-distribution problem is set up as follows: $\boldsymbol{X} = (X_1, \ldots, X_n)$ are jointly continuous with density function $f_{\boldsymbol{X}}$ over $\mathbb{R}^n$. We also have a measurable function $g : \mathbb{R}^n \to \mathbb{R}^n$ and we define the random variable $\boldsymbol{Y} = g(\boldsymbol{X})$. Our goal is to find a good means of finding the distribution of $\boldsymbol{Y}$ in terms of the distribution of $\boldsymbol{X}$.

In particular, we will make an assumption about g which is "well-behaved" in a few ways – and allows us to use the *Jacobian formula.* We will assume the following:

Assumption 0.1. Let $U \subset \mathbb{R}^n$ be an open set, and let $g : U \to \mathbb{R}^n$ be

- continuously differentiable
- an injection; and
- has non-vanishing determinant of the Jacobian, i.e. $\left|\frac{\partial g}{\partial \boldsymbol{x}}\right| \neq 0$.

We also have the following fact, which is super useful:

Fact 0.1. Define V as the image $g(U)$. Then if $g : U \to V$ satisfies the assumption: (i) V is open; (ii) $g^{-1} : V \to U$ is well-defined; (iii) g^{-1} satisfies the assumption as well.

Let us define $\boldsymbol{J}(\boldsymbol{y})$ to be the *Jacobian* (first-derivative, basically) of g^{-1} at $\boldsymbol{y}$. Basically, around any point $\boldsymbol{y}$, we consider a tiny cube A of volume δ^n and note that the probability mass inside came from the parallelepiped $B = g^{-1}(A) \approx \boldsymbol{J}(\boldsymbol{y})A$. The volume of it is then $\approx |\boldsymbol{J}(\boldsymbol{y})|\delta^n$ (linear algebra fact), and the density inside is approximately $f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y}))$. Thus, the mass ($\sim \boldsymbol{Y}$) inside A should be equal to the mass ($\sim \boldsymbol{X}$) in B, giving:

$$f_{\boldsymbol{y}}(\boldsymbol{y}) \cdot \delta^n \approx f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y})) \cdot |\boldsymbol{J}(\boldsymbol{y})|\delta^n$$

Dividing both sides by δ^n and then taking $\delta \searrow 0$ (which turns the $\approx$ into $=$), we get the actual **Jacobian formula**:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(g^{-1}(\boldsymbol{y})) \cdot |\boldsymbol{J}(\boldsymbol{y})|$$

For convenience, we will also be using the matrix $\boldsymbol{M} := \frac{\partial g}{\partial \boldsymbol{x}}(g^{-1}(\boldsymbol{y}))$ (forward Jacobian of g measured at $\boldsymbol{x} = g^{-1}(\boldsymbol{y})$). We will use the fact that $|\boldsymbol{J}(\boldsymbol{y})| = |\boldsymbol{M}|^{-1}$.

An innocent little problem using the Jacobian formula

Problem 0.1. *Let* $\boldsymbol{X} = (X_1, X_2)$ *be jointly continuous with PDF* $f_{\boldsymbol{X}}(x_1, x_2) = \exp(-x_1 - x_2)$ *for* $x_1, x_2 > 0$*, and let*

$$\boldsymbol{Y} = (Y_1, Y_2) = (X_1 + X_2, X_1 X_2)$$

We want to know: (a) what is the joint PDF of $\boldsymbol{Y}$*, and (b) are* Y_1, Y_2 *independent?*

Well, to (b) we can already answer "`no`" because if $Y_2 \geq 100$, then Y_1 has to be bigger than 1 and that basically settles it.

$$\text{(Formally, we say } \mathbb{P}[(Y_2 \geq 100) \cap (Y_1 \leq 1)] = 0 \neq \mathbb{P}[Y_2 \geq 100] \cdot \mathbb{P}[Y_1 \leq 1]\text{)}$$

But let's do this in the principled way.

First, we have an issue that g is not one-to-one (note that $g(x_1, x_2) = g(x_2, x_1)$); we will solve this by means of *order statistics*. We can assume that $x_1 \neq x_2$ because $\{\boldsymbol{x} : x_1 = x_2\}$ has Lebesgue measure 0. Define:

$$Z_1 = \min(X_1, X_2) \text{ and } Z_2 = \max(X_1, X_2)$$

From the order-statistics problem in the homework, we know that the PDF $f_{\boldsymbol{Z}}$ is

$$f_{\boldsymbol{Z}}(z_1, z_2) = \begin{cases} 2\exp(-z_1 - z_2) & \text{if } 0 < z_1 < z_2 \\ 0 & \text{otherwise} \end{cases}$$

Note here that our set $U \subset \mathbb{R}^2$ is now

$$U = \{\boldsymbol{z} : 0 < z_1 < z_2\}$$

which is indeed open, and g remains the same and is therefore still continuously differentiable. Finally, if we look at the Jacobian of g, we find that

$$\frac{\partial g}{\partial \boldsymbol{z}} = \begin{bmatrix} 1 & 1 \\ z_2 & z_1 \end{bmatrix} \text{ and so } \left|\frac{\partial g}{\partial \boldsymbol{z}}\right| = z_2 - z_1$$

whose determinant is not 0 since $z_2 \neq z_1$.

Ok, let's take a deep breath and remind ourselves of the Jacobian formula:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{Z}}(g^{-1}(\boldsymbol{y}))\, |\boldsymbol{J}(\boldsymbol{y})|$$

(hidden is a $\mathbf{1}_V(\boldsymbol{y})$ term, i.e. this only works on the range of g). We'll need to find these two parts, $f_{\boldsymbol{Z}}(g^{-1}(\boldsymbol{y}))$ and $|\boldsymbol{J}(\boldsymbol{y})|$.

<u>The Density at the Inverse:</u> This luckily turns out to be quite easy, since by definition $z_1 + z_2 = y_1$ when $\boldsymbol{y} = g(\boldsymbol{z})$. Therefore, the density can just be computed:

$$f_{\boldsymbol{Z}}(g^{-1}(\boldsymbol{y})) = 2\exp(-y_1)$$

<u>The Determinant:</u> For this, we gotta look at g^{-1}. Given $\boldsymbol{y}$, what is $\boldsymbol{z}$? Well, solving gives

$$y_2 = z_1(y_1 - z_1) = z_2(y_1 - z_2)$$

which can be solved quadratically. z_2 is the max, so

$$z_1 = \frac{y_1 - \sqrt{y_1^2 - 4y_2}}{2} \quad \text{and} \quad z_2 = \frac{y_1 + \sqrt{y_1^2 - 4y_2}}{2}$$

As a bit of a sanity check, let's look at $y_1^2 - 4y_2$, and hope that it's positive. We know

$$y_1^2 - 4y_2 = (x_1 + x_2)^2 - 4x_1x_2 \geq 0 \text{ because it's the square of AM-GM}$$

So our receiving set V is just

$$V = \{\boldsymbol{y} : y_1^2 - 4y_2 \geq 0\}$$

Alright, enough putting it off: what about the Jacobian $\boldsymbol{J}(\boldsymbol{y})$ of g^{-1}? To make things super-simple, however, note that we already have the determinant of the matrix $\boldsymbol{M}$, which is $z_1 - z_2$ (the *absolute value* of $\det(\boldsymbol{M})$ (at $\boldsymbol{z}$) is $z_2 - z_1$); and we know z_1 and z_2 in terms of y_1 and y_2. Thus, we get

$$\det(\boldsymbol{M}) = z_1 - z_2 = \frac{y_1 - \sqrt{y_1^2 - 4y_2}}{2} - \frac{y_1 + \sqrt{y_1^2 - 4y_2}}{2} = \sqrt{y_1^2 - 4y_2}$$

and therefore

$$\det(\boldsymbol{J}(\boldsymbol{y})) = \det(\boldsymbol{M})^{-1} = -\frac{1}{\sqrt{y_1^2 - 4y_2}}$$

Now, we take the absolute value of this to get what we needed:

$$|\boldsymbol{J}(\boldsymbol{y})| = \frac{1}{\sqrt{y_1^2 - 4y_2}}$$

Finally, we can put everything together that we needed – not forgetting the term that we hid (indicator of V) – to get

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{Z}}(g^{-1}(\boldsymbol{y}))\,|\boldsymbol{J}(\boldsymbol{y})|\,\mathbf{1}_V(\boldsymbol{y}) = \frac{2\exp(-y_1)}{\sqrt{y_1^2 - 4y_2}}\,\mathbf{1}_{\{y_1^2 - 4y_2 > 0\}}$$

As an afterthought, we get part (b) – are they indepedent? – is "no" (as we already knew) because this PDF does not factor nicely into a y_1 term and a y_2 term.

Conditional probability example

Problem 0.2. *Alice sends a bit to Bob; this is some $X \in \{-1, 1\}$, and the probability of $X = -1$ or 1 is 1/2 for each. However, the communication channel is noisy - in particular, it introduces some Gaussian noise $N \sim \mathcal{N}(0, 1)$ (which is independent from the transmitted bit). Bob then receives $Y = X + N$, and wants to remove the noise and recover the original bit.*

Bob finds that $Y = y$, for some $y \in \mathbb{R}$. Compute the probability $\mathbb{P}[X = 1 \,|\, Y = y]$.

This is a problem about conditioning with probability densities. Let $f_{Y|X}$ be the conditional density of Y given X, and let f_Y be the marginal density of Y. In this problem we want something of the form $\mathbb{P}[X|Y]$ but are really given things of the form $\mathbb{P}[Y|X]$ (and $\mathbb{P}[X]$) – so a natural approach is to use Bayes' formula.

Defining p_X to be the *probability mass function* of X, we get

$$\mathbb{P}[X = 1 \,|\, Y = y] = \frac{p_X(1) \cdot f_{Y|X}(y \,|\, 1)}{f_Y(y)}$$

Note that because the noise is $\mathcal{N}(0, 1)$ (and independent of X), note that $Y \sim \mathcal{N}(X, 1)$ for whatever X is. Therefore, the density

$$f_{Y|X}(y \,|\, x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}}$$

Furthermore, f_Y is built as an average of these (recalling that X can only take two values):

$$f_Y(y) = \sum_x p_X(x) \cdot f_{Y|X}(y \,|\, x) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(y+1)^2}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}}}{2} = \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{(y+1)^2}{2}} + e^{-\frac{(y-1)^2}{2}}}{2}$$

because $p_X(x) = 1/2$ for $x = -1, 1$. Plugging in all of these into the formula above yields (after a bunch of cancellations with the 1/2 and the $1/\sqrt{2\pi}$):

$$\mathbb{P}[X = 1 \,|\, Y = y] = \frac{p_X(1) \cdot f_{Y|X}(y \,|\, 1)}{f_Y(y)} = \frac{e^{-\frac{(y-1)^2}{2}}}{e^{-\frac{(y+1)^2}{2}} + e^{-\frac{(y-1)^2}{2}}} = \frac{e^y}{e^{-y} + e^y}$$

(the last step is just an algebraic simplification, cancelling out the $e^{-\frac{y^2+1}{2}}$ on the top and bottom).

Notably, this function has the following natural properties for this problem (sanity check):

- $\lim_{y \to -\infty} \mathbb{P}[X = 1 \,|\, Y = y] = 0$ and $\lim_{y \to \infty} \mathbb{P}[X = 1 \,|\, Y = y] = 1$.
- $\mathbb{P}[X = 1 \,|\, Y = y]$ is (strictly) monotonically increasing.
- $\mathbb{P}[X = 1 \,|\, Y = 0] = 1/2$.

Borel-Cantelli example

Problem 0.3. *Suppose we have a sequence of nonnegative random variables X_n (not necessarily independent) such that for any constant $c > 0$, the following holds:*

$$0 < \mathbb{P}[X_n > c] \leq \frac{1}{c^2}$$

We want to show the following two things:

- *(a) For any constant $b > 0$, there is 0 probability that* $\limsup_{n\to\infty} \frac{X_n}{n} > b$.
- *(b) With probability 1,* $\lim_{n\to\infty} \frac{X_n}{n} = 0$.

For part (a), this is all about getting the thing we want to prove into a format where we can hit it with the given inequality. Furthermore, recall that lim sup is basically an "infinitely often" thing, which suggests that we might want to apply *Borel-Cantelli*. This means:

$$\limsup_{n\to\infty} \frac{X_n}{n} > b \iff \left\{\frac{X_n}{n} > b \text{ i.o.}\right\}$$

(CAUTION! Need to be careful about the inequalities - if it's $\geq$ it becomes more complicated, see Grading Notes 1 and 3.) Furthermore, we can re-write it to make the given inequality applicable. Define:

$$A_n := \left\{\omega : \frac{X_n(\omega)}{n} > b\right\} = \{\omega : X_n(\omega) \geq bn$$

Then, applying the inequality, we get

$$\mathbb{P}[A_n] = \mathbb{P}[X_n > bn] \leq \frac{1}{b^2n^2}$$

Therefore, summing up these probabilities gives, for any $b > 0$,

$$\sum_n \mathbb{P}[A_n] = \sum_n \frac{1}{b^2n^2} = \frac{\pi^2}{9b^2} < \infty$$

Therefore, we can apply Borel-Cantelli to conclude that $\limsup_{n\to\infty} \frac{X_n}{n} > b$ has probability 0.

For part (b), there are two options available (both basically the same concept). First, note that because X_n is *nonnegative*, we know that $0 \leq \liminf_{n\to\infty} X_n \leq \limsup_{n\to\infty} X_n$. Therefore, if $\limsup_{n\to\infty} X_n = 0$, we know that $\limsup_{n\to\infty} X_n = 0 = \liminf_{n\to\infty} X_n$, and therefore $\lim_{n\to\infty} X_n$ exists and is 0. Thus,

$$\lim_{n\to\infty} \frac{X_n}{n} = 0 \iff \limsup_{n\to\infty} \frac{X_n}{n} = 0$$

So now we really need to write "$\limsup_{n\to\infty} X_n = 0$" (as an event) in terms of events we already have - and a countable number of them too. Defining

$$C := \left\{\omega : \limsup_{n\to\infty} \frac{X_n(\omega)}{n} = 0\right\} \text{ and } C_k := \left\{\omega : \limsup_{n\to\infty} \frac{X_n(\omega)}{n} \leq \frac{1}{k}\right\}$$

We then just see that (by the *union bound*, and part (a))

$$C = \bigcap_k C_k \implies C^c = \bigcup_k C_k^c \implies \mathbb{P}[C^c] \le \sum_k \mathbb{P}[C_k^c]$$

$$= \sum_k 0 = 0 \implies \mathbb{P}[C] = 1 - \mathbb{P}[C^c] = 1$$

Alternately, it can be observed that $C_k \searrow C$, and $\mathbb{P}[C_k] = 1$ for all k; therefore, by continuity of probability we can conclude that $\mathbb{P}[C] = \lim_{k\to\infty} \mathbb{P}[C_k] = 1$.