Mathematics Department Stanford University
Math 51H – Chain rule

As a warm up to the chain rule, let's talk about the composition of continuous functions.

Theorem 1 *Suppose (X, d_X), (Y, d_Y), (Z, d_Z) are metric spaces, $f : X \to Y$ is continuous at $a \in X$, $g : Y \to Z$ is continuous at $f(a) \in Y$. Then $g \circ f : X \to Z$ is continuous at a.*

Proof: Given $\varepsilon > 0$ take $\delta' > 0$ using the definition of continuity of g at $f(a)$ so $d_Y(y, f(a)) < \delta'$ implies $d_Z(g(y), g(f(a))) < \varepsilon$, and then take $\delta > 0$ using the definition of continuity of f at a for δ', so $d_X(x, a) < \delta$ implies $d_Y(f(x), f(a)) < \delta'$. Then $d_X(x, a) < \delta$ implies $d_Z(g(f(x)), g(f(a)) < \varepsilon$, giving the desired continuity. □

We are now ready for the chain rule.

Theorem 2 *Suppose $U \subset \mathbb{R}^n$, $V \subset \mathbb{R}^m$ are open, $f : U \to V$, $g : V \to \mathbb{R}^p$, $x \in U$, f differentiable at x, g differentiable at $f(x)$. Then $g \circ f$ is differentiable at x with*

$$(D(g \circ f))(x) = (Dg)(f(x))(Df)(x).$$

Proof: Write

$$f(x + h) = f(x) + (Df)(x)h + R_f(x, h),$$

i.e. define $R_f(x, h) = f(x + h) - (f(x) + (Df)(x)h)$, so the differentiability of f at x is equivalent to: for all $\varepsilon_f > 0$ there is $\delta_f = \delta_f(\varepsilon_f) > 0$ such that

$$\|h\| < \delta_f \Rightarrow \|R_f(x, h)\| \leq \varepsilon_f \|h\|.$$

Similarly, write

$$g(y + k) = g(y) + (Dg)(y)k + R_g(y, k)$$

so that the differentiability of g at $f(x)$ is equivalent to: for all $\varepsilon_g > 0$ there is $\delta_g = \delta_g(\varepsilon_g) > 0$ such that

$$\|k\| < \delta \Rightarrow \|R_g(f(x), k)\| \leq \varepsilon_g \|k\|.$$

In order to prove the theorem we need to show that for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\|h\| < \delta \Rightarrow \|g(f(x + h)) - g(f(x)) - (Dg)(f(x))(Df)(x)h\| \leq \varepsilon \|h\|. \tag{1}$$

So let us first express $g(f(x + h))$ using the notation we introduced. We have

$$g(f(x + h)) = g(f(x) + (Df)(x)h + R_f(x, h)),$$

so with $k = (Df)(x)h + R_f(x, h)$ we get

$$g(f(x + h)) = g(f(x)) + (Dg)(f(x))((Df)(x)h + R_f(x, h)) + R_g(f(x), (Df)(x)h + R_f(x, h)),$$

and so

$$\begin{aligned} &\|g(f(x + h)) - g(f(x)) - (Dg)(f(x))(Df)(x)h\| \\ &= \|(Dg)(f(x))R_f(x, h) + R_g(f(x), (Df)(x)h + R_f(x, h))\| \\ &\leq \|(Dg)(f(x))R_f(x, h)\| + \|R_g(f(x), (Df)(x)h + R_f(x, h))\| \end{aligned}$$

by the triangle inequality. Thus, (1) is reached if given $\varepsilon > 0$ we find $\delta > 0$ such that

$$\|h\| < \delta \Rightarrow \|(Dg)(f(x))R_f(x, h)\| \leq \frac{\varepsilon}{2}\|h\| \text{ and } \|R_g(f(x), (Df)(x)h + R_f(x, h))\| \leq \frac{\varepsilon}{2}\|h\|. \tag{2}$$

Now, the first inequality is easy to arrange: as

$$\|(Dg)(f(x))R_f(x, h)\| \leq \|(Dg)(f(x))\| \|R_f(x, h)\|,$$

it suffices if we arrange

$$\|R_f(x,h)\| \le \frac{\varepsilon}{2(\|(Dg)(f(x))\|+1)}\|h\|$$

for then

$$\|(Dg)(f(x))R_f(x,h)\| \le \|(Dg)(f(x))\|\|R_f(x,h)\| \le \frac{\varepsilon\|(Dg)(f(x))\|}{2(\|(Dg)(f(x))\|+1)}\|h\| \le \frac{\varepsilon}{2}\|h\|.$$

But this is now easy: apply the definition of differentiability of f with $\varepsilon_f = \frac{\varepsilon}{2(\|(Dg)(f(x))\|+1)}$ to get

$$\delta_f = \delta_f\Big(\frac{\varepsilon}{2(\|(Dg)(f(x))\|+1)}\Big);$$

if we take any $\delta \le \delta_f$, then $\|h\| \le \delta$ implies $\|h\| \le \delta_f$ and thus that (2) holds.

We now turn to the second, more subtle inequality. By the definition of the differentiability of g at $f(x)$, we have that for any $\varepsilon_g > 0$ there is $\delta_g > 0$ such that

$$\|(Df)(x)h + R_f(x,h)\| < \delta_g \Rightarrow \|R_g(f(x),(Df)(x)h + R_f(x,h))\| \le \varepsilon_g\|(Df)(x)h + R_f(x,h)\|. \quad (3)$$

So clearly it is important to control $\|(Df)(x)h + R_f(x,h)\|$. Here $(Df)(x)h$ has size comparable to h, $R_f(x,h)$ can be made smaller than any multiple of h, but as we are adding this to $(Df)(x)h$ it makes no difference if we make the multiple small (the sum will not be a small multiple anyway). So let's use the definition of the differentiability of f at x with $\varepsilon_f = 1$: there exists $\delta_f = \delta_f(1)$ such that

$$\|h\| < \delta_f(1) \Rightarrow \|R_f(x,h)\| \le \|h\|.$$

Thus, for $\|h\| < \delta_f(1)$,

$$\|(Df)(x)h + R_f(x,h)\| \le \|Df(x)\|\,\|h\| + \|R_f(x,h)\| \le (\|Df(x)\|+1)\|h\|.$$

So, with ε_g to be determined still, we have

$$\begin{aligned} &\|h\| < \delta_f(1) \text{ and } (\|Df(x)\|+1)\|h\| < \delta_g(\varepsilon_g) \Rightarrow \\ &\qquad \|R_g(f(x),(Df)(x)h + R_f(x,h))\| \le \varepsilon_g\|(Df)(x)h + R_f(x,h)\| \le \varepsilon_g(\|Df(x)\|+1)\|h\| \end{aligned} \quad (4)$$

since $(\|Df(x)\|+1)\|h\| < \delta_g(\varepsilon_g)$ implies $\|(Df)(x)h + R_f(x,h)\| \le (\|Df(x)\|+1)\|h\| < \delta_g(\varepsilon_g)$, and now we apply (3). We are now very close. Let $\varepsilon_g = \frac{\varepsilon}{2(\|Df(x)\|+1)}$ to get

$$\delta_g = \delta_g\Big(\frac{\varepsilon}{2(\|Df(x)\|+1)}\Big).$$

Then by (4)

$$\|h\| < \delta_f(1) \text{ and } (\|Df(x)\|+1)\|h\| < \delta_g\Big(\frac{\varepsilon}{2(\|Df(x)\|+1)}\Big)$$

imply

$$\|R_g(f(x),(Df)(x)h + R_f(x,h))\| \le \frac{\varepsilon}{2(\|Df(x)\|+1)}(\|Df(x)\|+1)\|h\| = \frac{\varepsilon}{2}\|h\|.$$

So if

$$\|h\| < \delta_f(1) \text{ and } \|h\| < (\|Df(x)\|+1)^{-1}\delta_g\Big(\frac{\varepsilon}{2(\|Df(x)\|+1)}\Big),$$

then

$$\|R_g(f(x),(Df)(x)h + R_f(x,h))\| \le \frac{\varepsilon}{2}\|h\|.$$

So we now simply let δ to be the minimum of the three constraints we have for $\|h\|$:

$$\delta = \min\Big(\delta_f\Big(\frac{\varepsilon}{2(\|(Dg)(f(x))\|+1)}\Big), \delta_f(1), (\|Df(x)\|+1)^{-1}\delta_g\Big(\frac{\varepsilon}{2(\|Df(x)\|+1)}\Big)\Big);$$

then $\|h\| < \delta$ implies that

$$\|(Dg)(f(x))R_f(x,h)\| \le \frac{\varepsilon}{2}\|h\| \text{ and } \|R_g(f(x),(Df)(x)h + R_f(x,h))\| \le \frac{\varepsilon}{2}\|h\|,$$

i.e. (2) has been shown. This proves (1) and completes the proof. □