Author: Shane Guan

Motivation:
I wanted a method to express the nth order derivative without needing to use index notation. For instance, the first "derivative" of $f(x) = xx^T$ is a 3d tensor, whose terms are

$$\frac{\partial [xx^T]_{i,j}}{\partial x_k} = 1_{k=i} x_j + 1_{k=j} x_i$$

It's hard for me to visualize what this 3d tensor looks like. Also, this 3d tensor notation is only so that we can write $\Delta f \approx \frac{\partial f}{\partial x} \Delta x$ with the change in input Δx on the right and the derivative on the left of the product. I think if we get rid of that restriction, then we can simply write something like $\Delta(xx^T) \approx (\Delta x)x^T + x(\Delta x)^T$ which is intuitively what the 3d tensor is telling us it's doing.

I also believe that the method I developed can be used to easily find the final simplification of the nth order term in the Taylor expansion.

1 A method for understanding higher order derivatives of vector input, matrix output functions

Let's assume the only operations in the function $f(x)$ are dot product, matrix-matrix, matrix-vector product, and scalar operations. Let $x \in \mathbb{R}^m$ be a vector of m components.

1.1 Understanding the first order derivative

We are effectively finding another function $d_x^1 f$ that takes as input (x, δ), where δ is some vector input of the same shape as x. I pronounce this as the 1st order del function of f with respect to x. This function has the property such that $\frac{\partial f}{\partial x_i} = \frac{\partial (d_x^1 f)}{\partial \delta_i}$.

To find $(d_x^1 f)(x, \delta)$, we use the following rules

Base Case:

$$f(x) = x \implies d_x^1 f(x, \delta) = \delta$$
$$f(x) = x^T \implies d_x^1 f(x, \delta) = \delta^T$$
$$f(x) = A \implies d_x^1 f(x, \delta) = 0$$

Linearity:

$$f(x) = AW(x) + M(x)B \implies d_x^1 f = A(d_x^1 W) + (d_x^1 M)B$$

Product (matrix-matrix or matrix scalar, or matrix vector):

$$f(x) = W(x)M(x) \implies d_x^1 f(x, \delta) = W(x)(d_x^1 M) + (d_x^1 W)M(x)$$

Chain rule:

$$f(x) = W(M(x)) \implies d_x^1 f(x, \delta) = d_M^1 W(M(x), d_x^1 M(x, \delta))$$

In words, you first find the del function of W wrt its input M, and plug in $d_x^1 M$ for δ.

Element-wise function (where $*$ is element-wise multiply):

$$[f(x)]_i = s(x_i) \implies d_x^1 f(x, \delta) = s'(x) * \delta$$

Examples:

$$(d_x^1 x^T x)(x,\delta) = x^T\delta + \delta^T x = 2\langle \delta, x\rangle$$
$$(d_x^1 x x^T)(x,\delta) = x\delta^T + \delta x^T$$
$$(d_x^1 A x x^T B)(x,\delta) = Ax\delta^T B^T + A\delta x^T B^T = A(x\delta^T + \delta x^T)B^T$$
$$[d_x^1 x \sin(x)^T](x,\delta) = x(\cos(x) * \delta)^T + \delta \sin(x)^T$$

1.1.1 Conjectures

I believe that, given any f that satisfies our assumptions in the beginning, then $d_x^1 f(x,\delta)$ will always be linear in δ.

Also, for a vector-valued function $f(x)$, the first order del function is just the Jacobian J times δ:

$$d_x^1 f(x,\delta) = J\delta$$

1.2 Understanding the nth order derivative

The spirit for the nth order del function is the same. This is a function $d_x^n f(x,\Delta)$, where Δ is the set of n vectors $\Delta = \{\Delta_i \;:\; i \in [n],\; \Delta_i \in \mathbb{R}^m\}$. The first order del function has Δ as the singleton vector δ.

This nth order del function $d_x^n f(x,\Delta)$ is such that $\frac{\partial^n f}{\prod_{i\in\mathcal{S}} \partial x_i} = \frac{\partial^n (d_x^n f)}{\prod_{i\in[n]} \partial \Delta_{i,S_i}}$, where $\mathcal{S}$ is some n element bag of the indices of the vector x (so Δ_{i,S_i} is the S_ith component of the vector Δ_i). Higher order del functions will be found using this following method (where $\delta_{new} \notin \Delta$):

$$d_x^{n+1} f(x, \delta_{new} \cup \Delta) = [d_x^1(d_x^n f)](x, \delta_{new})$$

In other words, we take the 1st order del function of $d_x^n f(x,\Delta)$ with respect to x (letting Δ be treated as a constant).

Example: Suppose

$$f(x) = Axx^T B$$

Then

$$d_x^1 f = A(x\Delta_1^T + \Delta_1 x^T)B$$
$$d_x^2 f = A(\Delta_2\Delta_1^T + \Delta_1\Delta_2^T)B$$
$$d_x^3 f = 0$$

1.2.1 What is it really doing?

I believe that if you replace every element of Δ with the same vector δ, then $d_x^n f(x,\Delta)$ is the same as the nth term in the Taylor expansion of $f(x+\delta)$ scaled by $n!$. This follows if the linearity conjecture in section 1.1.1 holds.

1.2.2 A more detailed example

Suppose you want to show that the following function is non-smooth as $\|x\|_2 \to 0$ (the function comes from ridge regression, and $\|w^*\|_2 = 1$)

$$f(x) = \left\| w^* - \frac{x}{\|x\|_2} \right\|_2^2$$

Or in other words that the spectral norm of f is unbounded by a constant as $\|x\|_2 \to 0$. The gradient is

$$g(x) = \nabla_x f = \frac{1}{\|x\|_2}(I - \frac{xx^T}{\|x\|_2^2})(\frac{x}{\|x\|_2} - w^*) = \frac{1}{\|x\|_2}(\frac{xx^T}{\|x\|_2^2} - I)w^*$$

Now we find the del function of each of the terms in the gradient

$$d_x^1 \frac{1}{\|x\|_2} = d_x^1[(x^Tx)^{\frac{-1}{2}}] = \frac{-1}{2}\|x\|_2^{-3}(2x^T\delta) = -\frac{x^T\delta}{\|x\|_2^3}$$

$$d_x^1 \frac{x}{\|x\|_2} = -\frac{xx^T\delta}{\|x\|_2^3} + \frac{\delta}{\|x\|_2} = \frac{1}{\|x\|_2}(I - \frac{xx^T}{\|x\|_2^2})\delta$$

$$d_x^1(\frac{xx^T}{\|x\|_2^2}) = \frac{1}{\|x\|_2^2}(d_x^1 xx^T) + xx^T(\frac{-1}{\|x\|_2^4}2x^T\delta) = \frac{1}{\|x\|_2^2}(\delta x^T + x\delta^T) + xx^T(\frac{-1}{\|x\|_2^4}2x^T\delta)$$

$$d_x^1(\frac{xx^T}{\|x\|_2^2}) = \frac{1}{\|x\|_2^2}[(I - \frac{2xx^T}{\|x\|_2^2})\delta x^T + x\delta^T]$$

Now we combine the terms to find something representing the Hessian

$$d_x^1 g = (d_x^1 \frac{1}{\|x\|_2})(\frac{xx^T}{\|x\|_2} - I)w^* + \frac{1}{\|x\|_2}[d_x^1(\frac{xx^T}{\|x\|_2} - I)]w^*$$

$$d_x^1 g = -\frac{x^T\delta}{\|x\|_2^3}(\frac{xx^T}{\|x\|_2} - I)w^* + \frac{1}{\|x\|_2^3}[(I - \frac{2xx^T}{\|x\|_2^2})\delta x^T + x\delta^T]w^*$$

$$d_x^1 g = \frac{1}{\|x\|_2^3}[(I - \frac{xx^T}{\|x\|_2})x^T\delta + (I - \frac{2xx^T}{\|x\|_2^2})\delta x^T + x\delta^T]w^*$$

Because the spectral norm of the Hessian is $\sup_{\|\delta\|_2=1} \|H\delta\|_2$, and $d_x^1 g = H\delta$ by definition, it is clear that bounding the spectral norm of the Hessian is the same as bounding $\sup_{\|\delta\|_2=1} \|d_x^1 g\|_2$. Since we are showing that the spectral norm is not bounded by a constant for any x as $\|x\|_2 \to 0$, we can take $\langle x, w^* \rangle = 0$. Also, suppose $\delta = \frac{x}{\|x\|_2}$. Then

$$d_x^1 g(x, \delta) = \frac{1}{\|x\|_2^3}[x^T\delta w^* + x\delta^T w^*] = \frac{1}{\|x\|_2^3}[x^T\delta w^*]$$

$$\|d_x^1 g\|_2^2 = \frac{1}{\|x\|_2^6}[(x^T\delta)^2\|w^*\|_2^2] = \frac{1}{\|x\|_2^4}$$

Hence

$$\sup_{\|\delta\|_2=1} \|d_x^1 g\|_2^2 \geq \frac{1}{\|x\|_2^4}$$

Thus we have shown that f is non-smooth as $\|x\|_2 \to 0$.