TMA4215 – Project 2 Supplement

October 16, 2020

The chain rule. Let us derive the gradient in component form of a simple composition $F(y) = G \circ \Phi(y) = G(\Phi(y))$ where $G : \mathbb{R}^d \to \mathbb{R}$ and where $\Phi : \mathbb{R}^d \to \mathbb{R}^d$, so that

$$\Phi(y) = [\Phi_1(y_1, \ldots, y_d), \ldots, \Phi_d(y_1, \ldots, y_d)]^T$$

If we write $z_i = \Phi_i(y_1, \ldots, y_d)$ we find

$$[\nabla F(y)]_i = \frac{\partial F}{\partial y_i}(y) = \sum_{j=1}^{d} \frac{\partial G}{\partial z_j}(z) \frac{\partial \Phi_j}{\partial y_i}(y)$$

The Jacobian matrix of Φ is denoted $D\Phi$ and it has elements $[D\Phi]_{ji} = \frac{\partial \Phi_j}{\partial y_i}$. Therefore

$$[\nabla F(y)]_i = \sum_{j=1}^{d} \frac{\partial \Phi_j}{\partial y_i}(y) [\nabla G(z)]_j = \sum_{j=1}^{d} [D\Phi]_{ji} [\nabla G(z)]_j.$$

We conclude that

$$\nabla F(y) = (D\Phi(y))^T \nabla G(z), \quad z = \Phi(y)$$

Our case. Suppose $F : \mathbb{R}^d \to R$ is a differentiable function for which the gradient is sought. It has the form

$$F(y) = G \circ \Phi_{K-1} \circ \Phi_{K-2} \circ \cdots \circ \Phi_0(y)$$

The maps $\Phi_k : \mathbb{R}^d \to \mathbb{R}^d$, $0 \leq k \leq K-1$ and $G : \mathbb{R}^d \to \mathbb{R}$. The form of these maps are

$$\Phi_k(y) = y + h\sigma(W_k y + b_k), \quad G(y) = \eta(w^T y + \mu) \tag{1}$$

$W_k \in \mathbb{R}^{d \times d}$, $b_k \in \mathbb{R}^d$, $w \in \mathbb{R}^d$, $\mu \in \mathbb{R}$, $h \in \mathbb{R}$ are all known parameters (constants). To make this into a recursive procedure, it may be easier to define the functions

$$\Psi_0 = \Phi_0, \ \Psi_k = \Phi_k \circ \Psi_{k-1} = \Phi_k \circ \cdots \circ \Phi_0, \ k = 1, \ldots, K-1$$

Compared to the notation in the project, one has $Z^{(k)} = \Psi_{k-1}(y), \ k = 1, \ldots K$.

Then we have $F(y) = G \circ \Psi_{K-1}(y) = G(\Psi_{K-1}(y))$. The gradient of $F(y)$ is then

$$\nabla F(y) = (D\Psi_{K-1}(y))^T \nabla G(Z^{(K)})$$

Now, since $Z^{(K)} = \Psi_{K-1}(y) = \Phi_{K-1}(Z^{(K-1)}) = \Phi_{K-1} \circ \Psi_{K-2}(y)$ we get

$$D\Psi_{K-1}(y) = D\Phi_{K-1}(Z^{(K-1)}) \cdot D\Psi_{K-2}(y)$$

where the $\cdot$ is matrix-matrix multiplication (Jacobian$\times$Jacobian). The transpose is

$$(D\Psi_{K-1}(y))^T = (D\Psi_{K-2}(y))^T \cdot \Big(D\Phi_{K-1}(Z^{(K-1)})\Big)^T$$

We can repeat this all the way down, and so we could write this as a pseudo-code

```
Compute all the  Z^(k) in a forward sweep
A =∇G(Z^(K))
for k in range(K,0,-1):
    A = (DΦ_{k-1}(Z^(k-1))^T A
```

What remains now, is just to compute $\nabla G(Z^{(K)})$ and $(D\Phi_{k-1})^T$ using (1). Let us get back to coordinates again

$$G(y) = \eta\big(\sum_{i=1}^{d} w_i y_i + \mu\big) \quad \Rightarrow \quad \frac{\partial G}{\partial y_j} = \eta'\big(\sum_{i=1}^{d} w_i y_i + \mu\big) w_j$$

This is just

$$\nabla G(y) = \eta(w^T y + \mu) w$$

so a scalar times a vector. As for $\Phi_k(y)$ we simplify by "hiding" the k-index (just to focus on one of them), setting $\Phi(y) = y + h\sigma(Wy + b)$. In component form

$$[\Phi(y)]_i = y_i + h\sigma\big(\sum_{j=1}^{d} W_{ij} y_j + b_i\big) \quad \Rightarrow \quad \frac{\partial \Phi_i}{\partial y_r} = \delta_{ir} + h\sigma'\big(\sum_{j=1}^{d} W_{ij} y_j + b_i\big) W_{ir}$$

Note from the for-loop in the pseudo-code that we always just need to compute, for a vector A, the expression $(D\Phi(y))^T A$ whose rth component is

$$\sum_{i=1}^{d} [D\Phi(y)]_{ir} A_i = A_r + h \sum_{i=1}^{d} \sigma'\big(\sum_{j=1}^{d} W_{ij} y_j + b_i\big) W_{ir} A_i$$

The last term is of the form $\sum_{i=1}^{d} W_{ir}(A_i B_i)$ where $B_i = h\sigma'\big(\sum_{j=1}^{d} W_{ij} y_j + b_i\big)$. Since $(A \odot B)_i = A_i B_i$ we arrive at the update

$$D\Phi(y)^T A = A + W^T \big(h\sigma'(Wy + b) \odot A\big)$$