RECONSTRUCTION OF MASS CONSISTENT WIND VECTOR FIELDS BY A DEEP LEARNING METHOD

DANIEL CERVANTES, MIGUEL ANGEL MORELES

Contents

1. Introduction 1
2. The Frèchet Derivative 1
3. Derivatives of Neural Networks 4
4. The Inverse Problem 6
5. A basic Deep Learning algorithm 7
The $L-1$ hidden layers in a neural network 7
The optimization problem 7
6. The Gradient 8
6.1. Quadratic cost 8
6.2. Divergence cost with Finite Differences 9
6.3. Divergence cost 9
The stochastic gradient 9
References 9

1. Introduction

2. The Frèchet Derivative

Definition. Let E, F be Banach spaces, $f : E \to F$ and $x \in E$. If $A \in L(E, F)$ satisfies

$$\frac{\|f(x+h) - f(x) - Ah\|}{\|h\|} \to 0 \text{ as } h \to 0,$$

then f is said to be *differentiable* at x, and A is called the *(Frèchet) derivative* of f at x, denoted by $Df(x) = A$.

Example Let $f : E \to F$ be a constant function, i.e. $f(x) = k$. Then, $Df(x) = 0$, the zero linear function.

Indeed,

$$\frac{\|f(x+h)-f(x)-Ah\|}{\|h\|} =$$

$$\frac{\|k-k-0h\|}{\|h\|} = 0.$$

Example Let $f : E \to F$ be a linear function. Then, $Df(x) = f$.
Indeed,

$$\frac{\|f(x+h)-f(x)-f(h)\|}{\|h\|} =$$

$$\frac{\|f(x)+f(h)-f(x)-f(h)\|}{\|h\|} = 0.$$

Lemma. If f is differentiable at x, then

$$Df(x)h = \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon h) - f(x)}{\varepsilon}.$$

The right had side is called the Gateàux derivative, $\delta f(x, h)$, of f at x in direction h.

Example Let M_n be the space of $n \times n$ matrices. Let $f : M_n \to M_n$ be given by

$$f(X) = X^2.$$

f is a polynomial, thus a C^∞ function.

$$\frac{(X+\varepsilon H)^2 - X^2}{\varepsilon} =$$

$$\frac{X^2+\varepsilon(XH+HX)+\varepsilon^2 H^2 - X^2}{\varepsilon} =$$

$$XH + HX + \varepsilon H^2.$$

Consequently

$$Df(X)H = XH + HX.$$

Proposition. Properties:

(1) f differentiable at $x \Rightarrow f$ continuous at x.
(2) $D(af + bg) = aDf + bDg$.
(3) (Chain rule) $D(g \circ f)(x) = D(g(f(x)) \circ Df(x)$.

Theorem (Mean Value Theorem). If $U \subset E$ is open, $f : U \to F$ differentiable, and if

$$\Gamma = \{(1-t)x + ty : 0 \le t \le 1\} \subset U,$$

then

$$\|f(x) - f(y)\| \leq \sup_{\xi \in \Gamma} \|Df(\xi)\| \ \|x - y\| .$$

Partial derivatives

By $E_1 \oplus E_2$ is meant the space $E_1 \times E_2$ with the usual linear structure.

If E_1, E_2 are Banach (Hilbert), then $E_1 \oplus E_2$ is also Banach (Hilbert).

Notation. (Partial derivatives) $f : E_1 \oplus E_2 \to F$

$$D_1 f(x_1, x_2) = D\left(f(\cdot, x_2)\right)(x_1) \in L\left(E_1, F\right)$$

$$D_2 f(x_1, x_2) = D\left(f(x_1, \cdot)\right)(x_2) \in L\left(E_2, F\right)$$

Proposition. If $f : E_1 \oplus E_2 \to F$ is differentiable at (x_1, x_2), then $D_j f(x_1, x_2)$, $j = 1, 2$ exist and

$$Df(x_1, x_2)\left(\xi_1, \xi_2\right) = D_1 f(x_1, x_2)\xi_1 + D_2 f(x_1, x_2)\xi_2. \tag{(*)}$$

Conversely, if $U \subset E_1 \oplus E_2$ is open, and if $D_1 f$, $D_2 f$ exist and are continuos on U, then Df exists and $(*)$ holds.

Notation. Let E, F_1, F_2 Banach spaces, let $f : E \to F_1$, $g : E \to F_2$. Denote

$$(f, g) : E \to F_1 \oplus F_2,$$

$$(f, g)(x) = \left(f(x), g(x)\right).$$

Lemma. If $f : E \to F_1$, $g : E \to F_2$, and if f, g are differentiable at x, the (f, g) is differentiable at x and

$$D(f, g)(x) = \left(Df(x), Dg(x)\right),$$

i.e.

$$D(f, g)(x)\xi = \left(Df(x)\xi, Dg(x)\xi\right).$$

Proposition. If $\beta : E_1 \oplus E_2 \to F$ is bilinear and continuos, then β is differentiable on $E_1 \oplus E_2$ and

$$D\beta(x_1, x_2)\left(\xi_1, \xi_2\right) = \beta(\xi_1, x_2) + \beta(x_1, \xi_2).$$

Theorem (Leibniz rule). Let $u : E \to F_1$, $v : E \to F_2$ be differentiable at $x \in E$. If $\beta : F_1 \oplus F_2 \to G$ is bilinear and continuous, and if $f : E \to G$ is defined by

$$f(y) = \beta\left(u(y), v(y)\right), \quad \text{for } y \in E,$$

then f is differentiable at x and

$$Df(x)\xi = \beta(Du(x)\xi, v(x)) + \beta(u(x), Dv(x)\xi).$$

Example Let E be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Let $f : E \to \mathbb{R}$, $f(x) = \frac{1}{2}\left\|x\right\|^2$.

Consider $\beta : E \oplus E \to \mathbb{R}$ given by $\beta\left(x, y\right) = \frac{1}{2}\left\langle x, y\right\rangle$, and $U : E \to E \oplus E$, given by $U = \langle Id, Id\rangle$, Id the identity matrix. Then

$$f = \beta \circ U.$$

Thus

$$\begin{aligned}
Df(x)\xi &= D\beta(U(x)) \circ DU(x)\xi. \\
&= D\beta(U(x)) \circ (DId(x)\xi, DId(x)\xi) \\
&= D\beta(U(x)) \circ (\xi, \xi) \\
&= D\beta(x, x) \circ (\xi, \xi) \\
&= \beta(\xi, x) + \beta(x, \xi) \\
&= \frac{1}{2}\left\langle \xi, x\right\rangle + \frac{1}{2}\left\langle x, \xi\right\rangle \\
&= \left\langle \xi, x\right\rangle.
\end{aligned}$$

3. Derivatives of Neural Networks

Let us start by introducing an activation function

$$\sigma : \mathbb{R} \to \mathbb{R},$$

in essence a smoothed version of the step function. A popular choice is

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

For $z \in \mathbb{R}^d$,

$$\sigma : \mathbb{R}^d \to \mathbb{R}^d,$$

is defined component wise

$$\sigma(z)_i = \sigma(z_i).$$

Its derivative is

$$D\sigma(z) = diag(\sigma'(z_i)).$$

Let us denote the underlying weights matrix space in layer l by M_l, and by $A_l \equiv \mathbb{R}^{d_l}$ the biases space. Thus $a^{[l]}$, $b^{[l]} \in A_l$.

Let

$$\mathbf{X} = \prod_{l=1}^{L} (M_l \times A_l) .$$

the space of weights and biases. A typical element is

$$\mathbf{P} = \prod_{l=1}^{L} \left(W^{[l]}, b^{[l]}\right) .$$

For $j = 1, 2, \ldots, L$, let us introduce the map

$$\Sigma_j : A_{j-1} \times \mathbf{X} \to A_j \times \mathbf{X}$$

$$(a^{[j-1]}, \mathbf{P}) \mapsto (a^{[j]}, \mathbf{P}).$$

Here

$$a^{[j]} = (\sigma \circ z^{[j]})(a^{[j-1]}, (W^{[j]}, b^{[j]})),$$

and

$$z^{[j]} = W^{[j]} a^{[j-1]} + b^{[j]},$$

The Fréchet derivative of Σ_j at $(a^{[j-1]}, \mathbf{P})$ in direction $(\alpha_{j-1}, \mathbf{H})$ is denoted by

$$D\Sigma_j(a^{[j-1]}, \mathbf{P})(\alpha_{j-1}, \mathbf{H}), \quad \mathbf{H} = \prod_{l=1}^{L} (H_l \times h_l) .$$

It is readily seen that

$$Dz^{[j]}(a^{[j-1]}, (W^{[j]}, b^{[j]}))(\alpha_{j-1}, (H_j, h_j)) = W^{[j]}\alpha_{j-1} + H_j a^{[j-1]} + h_j.$$

Consequently

$$D\Sigma_j(a^{[j-1]}, \mathbf{P})(\alpha_{j-1}, \mathbf{H}) =$$

$$\left(Da^{[j]}(a^{[j-1]}, (W^{[j]}, b^{[j]}))(\alpha_{j-1}, (H_j, h_j)), \mathbf{H}\right) =$$

$$\left(diag\left(\sigma'(z_i^{[j]})\right)(W^{[j]}\alpha_{j-1} + H_j a^{[j-1]} + h_j), \mathbf{H}\right) .$$

Denote

$$\Delta_j = diag\left(\sigma'(z_i^{[j]})\right) .$$

For $j = 1, 2, \ldots, L$ define

$$\alpha_j = \Delta_j \left(W^{[j]}\alpha_{j-1} + H_j a^{[j-1]} + h_j\right) .$$

Consequently we mat write

$$D\Sigma_j(a^{[j-1]}, \mathbf{P})(\alpha_{j-1}, \mathbf{H}) = (\alpha_j, \mathbf{H}).$$

By the chain rule to $\Sigma_{j+1} \circ \Sigma_j$.

$$D(\Sigma_{j+1} \circ \Sigma_j)(a^{[j-1]}, \mathbf{P})(\alpha_{j-1}, \mathbf{H}) = (\alpha_{j+1}, \mathbf{H}) .$$

We are led to the following.

Theorem. Consider the projection map

$$\Pi_1 : A_L \times \mathbf{X} \to A_L,$$

and let

$$a_L : A_0 \times \mathbf{X} \to A_L,$$

be given by

$$a_L(a^{[0]}, \mathbf{P}) = \Pi_1 \circ \Sigma_L \circ \Sigma_{L-1} \circ \ldots \circ \Sigma_2 \circ \Sigma_1)(a^{[0]}, \mathbf{P}).$$

Then

$$Da_L(a^{[0]}, \mathbf{P})(\alpha_0, \mathbf{H}) = \alpha^{[L]}.$$

Moreover,

$$\begin{aligned}
&\alpha^{[L]} = \\
&(\Delta_L W_L)(\Delta_{L-1} W_{L-1}) \cdots (\Delta_2 W_2)(\Delta_1 W_1)\alpha_0 + \\
&\textstyle\sum_{l=1}^{L-1} (\Delta_L W_L) \cdots (\Delta_{l+1} W_{l+1}) \Delta_l (H_l a^{[l-1]} + h_l) + \\
&\Delta_L (H_L a^{[L-1]} + h_L)
\end{aligned}$$

Corollary.

$$D_{a^{[0]}} a_L(a^{[0]}, \mathbf{P})\alpha_0 = (\Delta_L W_L)(\Delta_{L-1} W_{L-1}) \cdots (\Delta_2 W_2)(\Delta_1 W_1)\alpha_0,$$

For $l = 1, 2, \ldots, L-1$,

$$D_{(W_l, b_l)} a_L(a^{[0]}, \mathbf{P})(H_l, h_l) = (\Delta_L W_L) \cdots (\Delta_{l+1} W_{l+1}) \Delta_l (H_l a^{[l-1]} + h_l),$$

and

$$D_{(W_L, b_L)} a_L(a^{[0]}, \mathbf{P})(H_L, h_L) = \Delta_L (H_L a^{[L-1]} + h_L).$$

Define

$$B_L = \Delta_L,$$

and

$$B_l = B_{l+1}(W_{l+1} \Delta_l), \quad l = 1, 2, \ldots, L-1.$$

We have

$$D_{(W_l, b_l)} a_L(a^{[0]}, \mathbf{P})(H_l, h_l) = B_l (H_l a^{[l-1]} + h_l), \quad l = 1, 2, \ldots, L.$$

4. The Inverse Problem

Assume that the first two components of a $3D$ vector field are known at N nonuniform points, namely, $\mathbf{U}_i^0 = (u_1(\mathbf{x}_i), u_2(\mathbf{x}_i))$, $i = 1, 2, \ldots, N \in \Omega \subset \mathbb{R}^3$.

Problem Construct a flow field $u(\mathbf{x} = (u_1(\mathbf{x}, u_2(\mathbf{x}, u_3(\mathbf{x})$ in the entire Ω, assuming that the set of discrete field values are known and that the approximant satisfies the continuity equation:

$$\nabla \cdot \mathbf{u} = 0.$$

$$J(\mathbf{u}) = \frac{1}{2}\|\mathcal{M}\mathbf{u} - \mathbf{U}^0\|^2.$$

We consider the problem:

Minimize $J(\mathbf{u})$ constrained to $\nabla \cdot \mathbf{u} = 0$.

5. A basic Deep Learning algorithm

We shall use the notation of Section 3.

The $L-1$ hidden layers in a neural network.

$$a^{[0]} = \mathbf{x} \in \Omega \subset A_0 = \mathbb{R}^3.$$

and

$$a^{[L]} = \mathbf{u} \in A_L = \mathbb{R}^3$$

The optimization problem. Let us denote the weights and bias parameters by,

$$(\mathbf{W}, \mathbf{b}) = (W^{[1]}, b^{[1]}, \ldots, (W^{[n+1]}, b^{[n+1]}).$$

At the ground level, the bottom part of the boundary of Ω, say Γ_b, we know that $u_3(\mathbf{x}) = 0$. Thus adding a a penalization term, we are led to minimize the cost function

$$\begin{aligned} L((\mathbf{P}) &= \textstyle\sum_{i=1}^N \frac{1}{2} [(\mathbf{u}_1(\mathbf{x}_i, \mathbf{P}) - \mathbf{U}_1^0)^2) + (\mathbf{u}_2(\mathbf{x}_i, \mathbf{P}) - \mathbf{U}_2^0)^2)] + \\ & \beta_1 \textstyle\int_\Omega |\nabla \cdot \mathbf{u}(\mathbf{x}; \mathbf{P})|^2 d\mathbf{x} + \beta_2 \int_{\Gamma_b} |u_3(\mathbf{x}; \mathbf{P})|^2 d\mathbf{x}_1 d\mathbf{x}_2 \end{aligned}$$

Notice that with the optimum, we obtain a field $\mathbf{u}(\mathbf{x})$.

Applying a simple quadrature

$$\begin{aligned} L((\mathbf{P}) &= \textstyle\sum_{i=1}^N \frac{1}{2} [(\mathbf{u}_1(\mathbf{x}_i, \mathbf{P}) - \mathbf{U}_1^0)^2) + (\mathbf{u}_2(\mathbf{x}_i, \mathbf{P}) - \mathbf{U}_2^0)^2)] + \\ & \beta_1 \textstyle\sum_{j=1}^m |\nabla \cdot \mathbf{u}(\mathbf{x}_j; \mathbf{P})|^2 \Delta \mathbf{x} + \beta_2 \sum_{k=1}^n |u_3(\mathbf{x}_k; \mathbf{P})|^2 \Delta \mathbf{S} \end{aligned}$$

Recall

$$D_{a^{[0]}} a_L(a^{[0]}, \mathbf{P})\alpha_0 = (\Delta_L W_L)(\Delta_{L-1} W_{L-1}) \cdots (\Delta_2 W_2)(\Delta_1 W_1)\alpha_0,$$

where

$$\Delta_l \equiv \Delta_l(a^{[0]}).$$

We can write

$$D_{a^{[0]}} a_L(a^{[0]}, \mathbf{P})\alpha_0 = B_1 W_1 \alpha_0,$$

It is readily seen that for the standard vectors $\mathbf{e}^j$, we have

$$\nabla \cdot \mathbf{u}(\mathbf{x}, \mathbf{P}) = B_1 W_1 \mathbf{e}^1 + B_1 W_1 \mathbf{e}^2 + B_1 W_1 \mathbf{e}^3.$$

Denoting $\mathbf{e} = \mathbf{e}^1 + \mathbf{e}^2 + \mathbf{e}^3$, we are led to consider the functional

$$\begin{aligned} L((\mathbf{P}) \quad = \quad & \textstyle\sum_{i=1}^{N} \frac{1}{2} [(\mathbf{u}_1(\mathbf{x}_i, \mathbf{P}) - \mathbf{U}_1^0)^2) + (\mathbf{u}_2(\mathbf{x}_i, \mathbf{P}) - \mathbf{U}_2^0)^2)] + \\ & \beta_1 \textstyle\sum_{j=1}^{m} |B_1(\mathbf{x}_j) W_1 \mathbf{e}|^2 \Delta \mathbf{x} + \beta_2 \textstyle\sum_{k=1}^{n} |u_3(\mathbf{x}_k; \mathbf{P})|^2 \Delta \mathbf{S} \end{aligned}$$

6. The Gradient

For any cost function

$$C : A_L \to \mathbb{R}.$$

By the chain rule we have

$$D(C \circ a_L)(a^{[0]}, \mathbf{P}) = DC(a_L(a^{[0]}, \mathbf{P}))\alpha^{[L]}.$$

6.1. **Quadratic cost.** Let Λ be a diagonal matrix.

$$C(a^{[L]}) = \frac{1}{2} \|\hat{y} - a^{[L]}\|_\Lambda^2 \equiv \frac{1}{2} \langle \Lambda(\hat{y} - a^{[L]}), \hat{y} - a^{[L]} \rangle,$$

Hence

$$D(C \circ a_L)(a^{[0]}, \mathbf{P}) = \langle \Lambda(a^{[L]} - \hat{y}), \alpha^{[L]} \rangle.$$

The gradient with respect to the bias b_l follows at once. Indeed,

$$D_{b_l}(C \circ a_L)(a^{[0]}, \mathbf{P}) h_l = \langle \Lambda(a^{[L]} - \hat{y}), B_l h_l \rangle.$$

hence

$$\nabla_{b_l}(C \circ a_L)(a^{[0]}, \mathbf{P}) = (B_l)^T \Lambda(a^{[L]} - \hat{y}).$$

On the other hand

$$D_{W_l}(C \circ a_L)(a^{[0]}, \mathbf{P}) H_l = \langle \Lambda(a^{[L]} - \hat{y}), B_l H_l a^{[l-1]} \rangle.$$

Given a matrix $\mathbf{Z}$, let us denote by $\mathbf{Z}_i$, and $\mathbf{Z}^j$ its $i-$th row and $j-$th column respectively.

We obtain

$$\begin{gathered} D_{W_l}(C \circ a_L)(a^{[0]}, \mathbf{P}) H_l = \\ \langle (B_l)^T \Lambda(a^{[L]} - \hat{y}), H_l a^{[l-1]} \rangle = \\ \textstyle\sum_i \left((B_l)^T \Lambda(a^{[L]} - \hat{y}) \right)_i \left(H_l a^{[l-1]} \right)_i = \\ \textstyle\sum_{i,j} \left((B_l)^T \Lambda(a^{[L]} - \hat{y}) \right)_i (H_l)_{i,j} (a^{[l-1]})_j. \end{gathered}$$

Consequently,

$$\nabla_{(W_l)_{i,j}}(C \circ a_L)(a^{[0]}, \mathbf{P}) = \left((B_l)^T \Lambda(a^{[L]} - \hat{y})\right)_i (a^{[l-1]})_j.$$

For the term

$$\left[\left(\mathbf{u}_1(\mathbf{x}_i, \mathbf{P}) - \mathbf{U}_1^0)^2\right) + \left(\mathbf{u}_2(\mathbf{x}_i, \mathbf{P}) - \mathbf{U}_2^0)^2\right)\right]$$

We let

$$\Lambda = diag(1, 1, 0), \quad \hat{y} = (\mathbf{U}_1^0, \mathbf{U}_2^0, 0),$$

whereas for the term

$$|u_3(\mathbf{x}_k; \mathbf{P})|^2,$$
$$\Lambda = diag(0, 0, 1), \quad \hat{y} = \mathbf{0}.$$

6.2. **Divergence cost with Finite Differences.** Let us start with the term

$$|\nabla \cdot \mathbf{u}(\mathbf{x}_j; \mathbf{P}), |^2$$

which can be approximated by

$$\left|\frac{\mathbf{u}(\mathbf{x}_{j,E}; \mathbf{P}) - \mathbf{u}(\mathbf{x}_{j,W}; \mathbf{P})}{2\Delta x} + \frac{\mathbf{u}(\mathbf{x}_{j,N}; \mathbf{P}) - \mathbf{u}(\mathbf{x}_{j,S}; \mathbf{P})}{2\Delta y} + \frac{\mathbf{u}(\mathbf{x}_{j,T}; \mathbf{P}) - \mathbf{u}(\mathbf{x}_{j,B}; \mathbf{P})}{2\Delta z}\right|^2$$

Define the cost function

$$C : (A_L)^6 \to \mathbb{R},$$

given by

$$C(a_E, a_W, a_N, a_S, a_T, a_B) = \left|\frac{a_E - a_W}{2\Delta x} + \frac{a_N - a_S}{2\Delta y} + \frac{a_T - a_B}{2\Delta z}\right|^2$$

$$DC(a_E, a_W, a_N, a_S, a_T, a_B)(\alpha_E, \alpha_W, \alpha_N, \alpha_S, \alpha_T, \alpha_B) =$$

$$\left\langle \tfrac{a_E - a_W}{2\Delta x} + \tfrac{a_N - a_S}{2\Delta y} + \tfrac{a_T - a_B}{2\Delta z}, \tfrac{\alpha_E - \alpha_W}{\Delta x} + \tfrac{\alpha_N - \alpha_S}{\Delta y} + \tfrac{\alpha_T - \alpha_B}{\Delta z} \right\rangle$$

6.3. **Divergence cost.** Now consider

$$C : \mathbf{M} \to \mathbb{R},$$

given by

$$C(B) = \frac{1}{2}|B\mathbf{e}|^2.$$

Then

$$DC(B)H = \langle B\mathbf{e}, H\mathbf{e} \rangle.$$

By the chain rule

$$D(C \circ B)(\mathbf{P})\mathbf{H} = DC(B(\mathbf{P}))DB(\mathbf{P})\mathbf{H},$$

or

$$D(C \circ B)(\mathbf{P})\mathbf{H} = \langle B(\mathbf{P})\mathbf{e}, DB(\mathbf{P})\mathbf{H}\mathbf{e} \rangle.$$

The stochastic gradient. See [1]

References

[1] Higham, C. F., & Higham, D. J. (2019). Deep learning: An introduction for applied mathematicians. Siam review, 61(4), 860 - 891.

[2] Weinan, E., & Yu, B. (2018). The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems. Communications in Mathematics and Statistics, 6(1), 1 - 12.