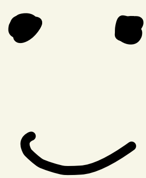
ORF 543
HW2

Evan Dogariu



(a) (5 points) Fix a sufficiently small positive number ϵ. Define $f_j(x)$ as in (1.2). Compute the absolute value of the jump in the slope of $f_j(x)$ at $x = \xi_j$, which is given by

$$\left| \frac{d}{dx} f_j(x) \Big|_{x=\xi_j+\epsilon} - \frac{d}{dx} f_j(x) \Big|_{x=\xi_j-\epsilon} \right| .$$

You answer should depend only $W_j^{(1)}, W_j^{(2)}$ (i.e. should not contain ϵ, ξ_j, σ, etc).

a) Fix $\varepsilon > 0$. There are two cases: $W_j^{(1)} < 0$ and $W_j^{(1)} > 0$

- $\underline{W_j^{(1)} < 0}$: In this case, $\forall x > \xi_j$ we have that $W_j^{(1)} x + b_j^{(1)} < 0$, and so $f_j'(x) = 0$. For $x \le \xi_j$, $W_j^{(1)} x + b_j^{(1)} \ge 0 \Rightarrow f_j(x) = W_j^{(2)}(W_j^{(1)} x + b_j^{(1)})$ and so $f_j'(x) = W_j^{(1)} W_j^{(2)}$. Then, $f_j'(\xi_j + \varepsilon) - f_j'(\xi_j - \varepsilon) = -W_j^{(1)} W_j^{(2)}$

- $\underline{W_j^{(1)} > 0}$: In this case, $\forall x \le \xi_j$ we have that $W_j^{(1)} x + b_j^{(1)} \le 0$, and so $f_j'(x) = 0$. For $x > \xi_j$, $W_j^{(1)} x + b_j^{(1)} > 0 \Rightarrow f_j(x) = W_j^{(2)}(W_j^{(1)} x + b_j^{(1)})$ and so $f_j'(x) = W_j^{(1)} W_j^{(2)}$. Then, $f_j'(\xi_j + \varepsilon) - f_j'(\xi_j - \varepsilon) = W_j^{(1)} W_j^{(2)}$

In either case, the magnitude of the jump is $\boxed{|W_j^{(1)} W_j^{(2)}|}$

$\square$

(b) (5 points) Show that there exist

$$\widetilde{W}_j^{(1)}, \widetilde{W}_j^{(2)}, \widetilde{b}_j^{(1)} \in \mathbb{R}$$

so that

$$f\left(x;\ W_j^{(1)}, b_j^{(1)}, W_j^{(2)}\right) = f\left(x;\ \widetilde{W}_j^{(1)}, \widetilde{b}_j^{(1)}, \widetilde{W}_j^{(2)}\right) \qquad \text{for all } x \in \mathbb{R}$$

and

$$\left|\widetilde{W}_j^{(1)}\right| = \left|\widetilde{W}_j^{(2)}\right|.$$

[Hint: It maybe useful to note that if c is any *non-negative* constant, then for any $t \in \mathbb{R}$ we have $\sigma(ct) = c\sigma(t)$.]

b) Define $c = \sqrt{\left|\frac{W_j^{(2)}}{W_j^{(1)}}\right|} > 0$ and $\widetilde{W}_j^{(1)} = c\,W_j^{(1)}$, $\widetilde{b}_j^{(1)} = c\,\widetilde{b}_j$

$\widetilde{W}_j^{(2)} = \frac{1}{c} W_j^{(2)}$

Then,

$$f_j(x;\ \widetilde{W}_j^{(1)}, \widetilde{W}_j^{(2)}, \widetilde{b}_j^{(1)}) = \widetilde{W}_j^{(2)} \sigma\left(c\left(W_j^{(1)}x + b_j^{(1)}\right)\right) \underset{\text{by the hint}}{=\!=} \frac{1}{c}\cdot c \cdot W_j^{(2)} \sigma\left(W_j^{(1)}x + b_j^{(1)}\right)$$

$$= f_j\left(x;\ W_j^{(1)}, W_j^{(2)}, b_j^{(1)}\right)$$

Also, $\left|\widetilde{W}_j^{(2)}\right| = \frac{1}{c}\cdot\left|W_j^{(2)}\right| = \sqrt{\left|\frac{W_j^{(1)}}{W_j^{(2)}}\right|}\cdot\left|W_j^{(2)}\right| = \sqrt{\left|W_j^{(1)}W_j^{(2)}\right|}$

and $\left|\widetilde{W}_j^{(1)}\right| = c\cdot\left|W_j^{(1)}\right| = \sqrt{\left|\frac{W_j^{(2)}}{W_j^{(1)}}\right|}\cdot\left|W_j^{(1)}\right| = \sqrt{\left|W_j^{(1)}W_j^{(2)}\right|}$

So, $\boxed{\left|\widetilde{W}_j^{(1)}\right| = \left|\widetilde{W}_j^{(2)}\right|}$ as desired.

$\square$

(c) (5 points) Use (b) to conclude that given

$$\theta = \left(b, b_j^{(1)}, W_j^{(1)}, W_j^{(2)}, \quad j = 1, \ldots, n\right), \qquad W_j^{(1)}, W_j^{(2)} \neq 0$$

there exists

$$\widetilde{\theta} = \left(\widetilde{b}, \widetilde{W}_j^{(1)}, \widetilde{W}_j^{(2)}, \widetilde{b}_j^{(1)}, \quad j = 1, \ldots, n\right), \qquad \widetilde{W}_j^{(1)}, \widetilde{W}_j^{(2)} \neq 0$$

so that

$$z(x; \theta, n) = z(x; \widetilde{\theta}, n), \qquad \text{for all } x \in \mathbb{R}$$

and

$$\frac{1}{2}\sum_{j=1}^{n}\left[\left(\widetilde{W}_j^{(1)}\right)^2 + \left(\widetilde{W}_j^{(2)}\right)^2\right] = \sum_{j=1}^{n}\left|\widetilde{W}_j^{(1)}\widetilde{W}_j^{(2)}\right|.$$

c) Define $\tilde{b} = b$ and for all j:

$$\tilde{W}_j^{(2)} = \sqrt{\left|\frac{W_j^{(1)}}{W_j^{(2)}}\right|}\, W_j^{(2)}, \quad \tilde{W}_j^{(1)} = \sqrt{\left|\frac{W_j^{(2)}}{W_j^{(1)}}\right|}\, W_j^{(1)}, \quad \tilde{b}_j^{(1)} = \sqrt{\left|\frac{W_j^{(2)}}{W_j^{(1)}}\right|}\, b_j^{(1)}$$

Part (b) yields that $f(x; W_j^{(1)}, W_j^{(2)}, b_j^{(1)}) = f(x; \tilde{W}_j^{(1)}, \tilde{W}_j^{(2)}, \tilde{b}_j^{(1)}) \quad \forall j$

$$\Rightarrow \boxed{z(x;\theta,n) = z(x;\tilde{\theta},n)} \quad \text{since} \quad z(x;\theta,n) = b + \sum_{j=1}^{n} f(x; W_j^{(1)}, W_j^{(2)}, b_j^{(1)}) = \tilde{b} + \sum_{j=1}^{n} f(x; \tilde{W}_j^{(1)}, \tilde{W}_j^{(2)}, \tilde{b}_j^{(1)})$$

Part (b) also gives that $|\tilde{W}_j^{(1)}| = |\tilde{W}_j^{(2)}| \quad \forall j$

$$\Rightarrow \frac{1}{2}\left(\tilde{W}_j^{(1)2} + \tilde{W}_j^{(2)2}\right) = \frac{1}{2}\left(|\tilde{W}_j^{(1)}|^2 + |\tilde{W}_j^{(2)}|^2\right)$$

$$= \frac{1}{2}\left(|\tilde{W}_j^{(1)}||\tilde{W}_j^{(2)}| + |\tilde{W}_j^{(1)}||\tilde{W}_j^{(2)}|\right)$$

$$= |\tilde{W}_j^{(1)}\tilde{W}_j^{(2)}| \qquad \forall j$$

So,

$$\boxed{\frac{1}{2}\sum_{j=1}^{n}\left[(\tilde{W}_j^{(1)})^2 + (\tilde{W}_j^{(2)})^2\right] = \sum_{j=1}^{n}|\tilde{W}_j^{(1)}\tilde{W}_j^{(2)}|}$$

D

(d) (5 points) Explain intuitively why the collection of functions $\mathcal{F}(\mathcal{D})$ can approximately be thought of as follows

Call this $\mathcal{F}_\lambda(\mathcal{D})$ $\rightarrow$

$$\mathcal{F}(\mathcal{D}) \approx \{z(x;\theta,n) \mid \mathcal{L}_{MSE}(\theta) + \lambda R(\theta) \text{ is minimal}, \quad n \text{ is arbitrary}\},$$

when λ is very small and where

$$\mathcal{L}_{MSE}(\theta) = \sum_{i=1}^{n_d} \left(y^{(i)} - z(x^{(i)};\theta)\right)^2.$$

d) Firstly, note that $\mathcal{L}_{MSE}(\theta)$ is minimized if and only if $z(x;\theta)$ interpolates the data. So, for every $z(x;\theta^*) \in \mathcal{F}(\mathcal{D})$, $\mathcal{L}_{MSE}(\theta^*) = 0$ $\Rightarrow \mathcal{L} = \lambda R(\theta^*)$. For very small λ we expect this to be minimal. More precisely, since $\mathcal{L} \geq 0$ always, for any other possible minimum $\mathcal{L}(\theta) = C = \mathcal{L}_{MSE}(\theta) + \lambda R(\theta)$, we can choose a λ small enough that

$$\lambda(R(\theta^*) - R(\theta)) < \mathcal{L}_{MSE}(\theta) \Rightarrow \mathcal{L}(\theta^*) < \mathcal{L}(\theta) \Rightarrow z(x;\theta^*) \in \mathcal{F}_\lambda(\mathcal{D}).$$

So, $\mathcal{F}(\mathcal{D}) \subseteq \mathcal{F}_\lambda(\mathcal{D})$ for small enough λ.

Now, note that when $\lambda = 0$, every $z(x;\theta^*) \in \mathcal{F}_\lambda(\mathcal{D})$ interpolates the data. We can increase λ by a small enough amount that any $z(x;\theta^*) \in \mathcal{F}_\lambda(\mathcal{D})$ still interpolates the data, but also has a nonzero $\lambda R(\theta^*)$ term. Therefore, any $z(x;\theta^*) \in \mathcal{F}_\lambda(\mathcal{D})$ will be a model that interpolates the data and has a smaller $R(\theta^*)$ than all other θ's whose $z(x;\theta)$ also interpolate the data $\Rightarrow z(x;\theta^*) \in \mathcal{F}(\mathcal{D})$. So, for small enough λ, $\mathcal{F}_\lambda(\mathcal{D}) \subseteq \mathcal{F}(\mathcal{D})$.

Therefore for small enough λ, $\mathcal{F}_\lambda(\mathcal{D}) = \mathcal{F}(\mathcal{D})$

and the result follows.

$\square$

(e) (5 points) Use (c) to obtain the following alternative description of $\mathcal{F}(\mathcal{D})$:

$$\mathcal{F}(\mathcal{D}) = \left\{ z(x;\theta,n) \mid z(x^{(i)};\theta,n) = y^{(i)} \text{ for all } i, \quad n \text{ is arbitrary}, \quad \widetilde{R}(\theta) \text{ is minimal} \right\},$$

call this $\tilde{\mathcal{F}}(\mathcal{D})$ for this problem

where

$$\widetilde{R}(\theta) = \sum_{j=1}^{n} \left| W_j^{(1)} W_j^{(2)} \right|$$

is a new regularizer.

e) Suppose that you have some θ^* such that $z(x;\theta^*)$ interpolates the data and

$$R(\theta^*) \leq R(\theta) \quad \forall \theta \quad \text{s.t.} \quad z(x;\theta) \text{ interpolates the data.}$$

Note that $\forall a,b \quad \frac{1}{2}(|a|-|b|)^2 \geq 0 \Leftrightarrow \frac{1}{2}(|a|^2+|b|^2) \geq |ab|$

So, $\tilde{R} \leq R$ always. Therefore,

$$(*) \quad \tilde{R}(\theta^*) \leq R(\theta) \quad \forall \theta \quad \text{s.t.} \quad z(x;\theta) \text{ interpolates the data.}$$

Now, by Part (c) we know that $\forall \theta \; \exists \tilde{\theta}$ s.t. $z(x;\tilde{\theta}) = z(x;\theta)$, $|\tilde{w}_j^{(1)}| = |\tilde{w}_j^{(2)}|$, and $\tilde{R}(\tilde{\theta}) = R(\tilde{\theta})$

Note that since $\tilde{w}_j^{(1)} = c\, w_j^{(1)}$ and $\tilde{w}_j^{(2)} = \frac{1}{c} w_j^{(2)}$, then $\tilde{w}_j^{(1)} \tilde{w}_j^{(2)} = w_j^{(1)} w_j^{(2)}$ in our construction of $\tilde{\theta}$ from θ. So, $\tilde{R}(\tilde{\theta}) = \tilde{R}(\theta)$

Together with $(*)$, this gives

$$\tilde{R}(\theta^*) \leq \tilde{R}(\theta) \quad \forall \theta \quad \text{s.t.} \quad z(x;\theta) \text{ interpolates the data}$$

$$\underline{\text{and}} \quad |w_j^{(1)}| = |w_j^{(2)}| \text{ for all weights in } \theta.$$

However, Part (c) gives that $\forall \theta_1$ that interpolate the data, there is some θ_2 that also interpolates the data with $\tilde{R}(\theta_1) = \tilde{R}(\theta_2)$ (our rescaling doesn't change $\tilde{R}$) and with θ_2 satisfying the equal weight magnitudes constraint.

So, $\forall \theta_1$ that interpolates the data, $\tilde{R}(\theta_1) = \tilde{R}(\theta_2) \geq \tilde{R}(\theta^*)$.

So, $\tilde{R}(\theta^*) \leq \tilde{R}(\theta) \quad \forall \theta$ s.t. $z(x;\theta)$ interpolates the data and therefore a minimum of R that interpolates the data also minimizes $\tilde{R}$. $\Rightarrow \mathcal{F}(\mathcal{D}) \subseteq \tilde{\mathcal{F}}(\mathcal{D})$

We can do the same thing for any θ^* s.t. $z(x;\theta^*)$ interpolates the data and $\tilde{R}(\theta^*) \leq \tilde{R}(\theta) \quad \forall \theta$ s.t. $z(x;\theta)$ interpolates.

Part (c) gives that $\exists\, \tilde{\theta}^*$ s.t. $\tilde{\theta}^*$ has equal magnitude weights $\Rightarrow \tilde{R}(\tilde{\theta}^*) = R(\tilde{\theta}^*)$ and also that $\tilde{R}(\tilde{\theta}^*) = \tilde{R}(\theta^*)$ since we scale each $W_j^{(1)*}$ by c and each $W_j^{(2)*}$ by $\frac{1}{c}$. This gives that

$$\tilde{R}(\theta^*) = \tilde{R}(\tilde{\theta}^*) = R(\tilde{\theta}^*) \leq \tilde{R}(\theta) \leq R(\theta) \quad \forall \theta \text{ s.t. } z(x;\theta) \text{ interpolates}$$

where the last inequality is because $\tilde{R} \leq R$ always. So, $\tilde{\theta}^*$ minimizes R with $z(x;\tilde{\theta}^*)$ interpolating the data. Since $z(x;\theta^*) = z(x;\tilde{\theta}^*)$, every function in $\tilde{\mathcal{F}}(D)$ has a form for some $\tilde{\theta}^*$ that also minimizes R. So, $\tilde{\mathcal{F}}(D) \subseteq \mathcal{F}(D)$.

Together, we get our result. $\square$

(f) (5 points) Fix n and consider $z(x;\theta,n)$. Under this assumption show that

$$\widetilde{R}(\theta)=\sum_{j=1}^{n}\left|s_{j+1}-s_j\right|.$$

f) Let $\xi_0=-\infty$, $\xi_{n+1}=\infty$ for notational purposes. For each $j=1,\dots,n$, let $x_-=\xi_j-\varepsilon\in(\xi_{j-1},\xi_j)$ and $x_+=\xi_j+\varepsilon\in(\xi_j,\xi_{j+1})$. Now, note that $\forall i\neq j$, $f_i'(x_-)=f_i'(x_+)$ since x_- and x_+ lie on the same side of the breakpoint ξ_i. So,

$$|s_{j+1}-s_j|=|z'(x_+;\theta,n)-z'(x_-;\theta,n)|=|f_j'(x_+)-f_j'(x_-)|$$

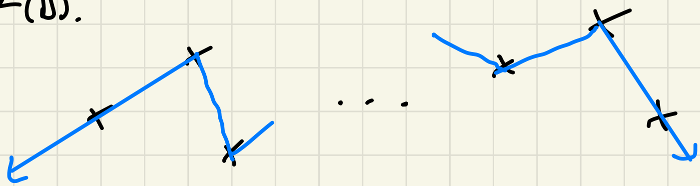
By part (a), we have that the magnitude of this derivative jump in f_j is precisely $|W_j^{(1)}W_j^{(2)}|$. So,

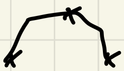
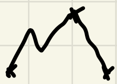
$$\sum_{j=1}^{n}|s_{j+1}-s_j|=\sum_{j=1}^{n}|W_j^{(1)}W_j^{(2)}|=\widetilde{R}(\theta).$$

$\square$

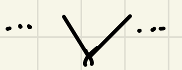
(g) For a fixed dataset as above, produce one example of a function $f \in \mathcal{F}(\mathcal{D})$.

g) We want to show that the piecewise linear function that connects all the datapoints and continues off to $\pm\infty$ as drawn below is in $\mathcal{F}(\mathcal{D})$.

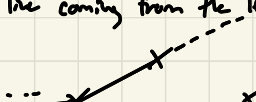


This clearly interpolates the data, and so we want to show that it has minimal slope jumps. Note that adding extra breakpoints can never decrease the sum of slope jumps: if we have , adding breakpoints like  keeps it the same, while adding breakpoints like  certainly makes things worse (this is always the case, even if the current breakpoints aren't at data points).
So, even though there are other elements of $\mathcal{F}(\mathcal{D})$ with more breakpoints, we know that our construction is a minimum of $\tilde{R}(\theta)$ with respect to the number of breakpoints.

Now, we want to show that moving breakpoints around cannot improve things. Intuitively, any time we have a breakpoint of the form , we would like to move the breakpoint down to flatten things.

However, it is already as low as it can be to fit the data, and so it must remain there. Similar logic holds for .

More precisely, for any triplet of consecutive data points, the breakpoint in the middle must lie on the dotted line coming from the left (assuming inductively that everything to the left is already optimal).



In order to fit the data, the next breakpoint must be on the dotted line and the next segment must go through the third data point.

Clearly, the slope jumps are minimized when this is placed at the second data point regardless of whether the third point is above or below the dotted line. We have seen that if it is optimal to place breakpoints at the data points up to some point i, it is optimal to continue by placing the next breakpoint at datapoint i. By induction, our construction has optimal $\tilde{R}$ w.r.t. breakpoint placing as well. Since # and location of breakpoints uniquely define a piecewise continuous linear function, our construction is $\in \mathcal{F}(D)$.

$\square$

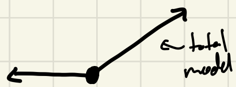
(i) There is a subtle flaw in our characterization of $\mathcal{F}(\mathcal{D})$! What is it? And how do you fix it?

i) There are several issues that manifest because of the potential for multiple neurons to have breakpoints at the same place. I list two below.

① If we have only one data point, we can use two neurons with breakpoints at that data point to model <u>any</u> line passing through the point. We can also model it with a single neuron. I write the two below for a point (x_1, y_1). Both models pass through the data point $\forall \varepsilon, \delta$.

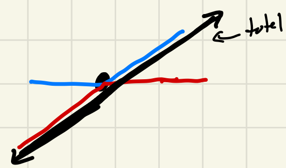
<u>one neuron</u>

$$z(x) = y_1 + \varepsilon\,\theta(\delta x - \delta x_1)$$



<u>two neurons</u>

$$z(x) = y_1 + \underbrace{\varepsilon\,\theta(\delta x - \delta x_1)}_{\text{neuron 1}} - \underbrace{\varepsilon\,\theta(-\delta x + \delta x_1)}_{\text{neuron 2}}$$

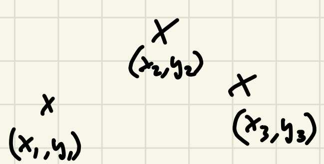


$\tilde{R} = \varepsilon\delta$, slope jump $= \varepsilon\delta$

$\tilde{R} = 2\varepsilon\delta$, slope jump $= 0$

So, the one neuron setting minimizes $\tilde{R}$ (and R), but the two neuron model minimizes the slope jump.

② Consider the three leftmost points of a general dataset.



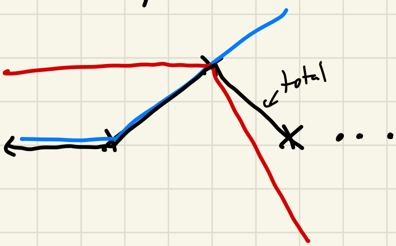
We can connect the first three points with at least two different two-neuron models:

model 1

$$z(x) = y_1 + \overbrace{\frac{y_2-y_1}{x_2-x_1}\theta(x-x_1)}^{\text{neuron 1}}$$

$$+ \underbrace{\left(\frac{y_3-y_2}{x_3-x_2} - \frac{x_3-x_1}{x_3-x_2}\cdot\frac{y_2-y_1}{x_2-x_1}\right)\theta(x-x_2)}_{\text{neuron 2}}$$

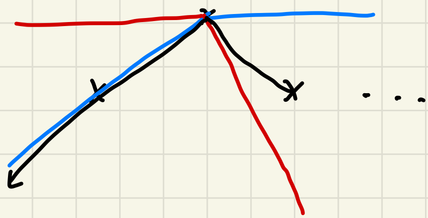
$$+ \dots$$



model 2

$$z(x) = y_2 - \overbrace{\frac{y_2-y_1}{x_2-x_1}\theta(-x+x_2)}^{\text{neuron 1}}$$

$$+ \underbrace{\frac{y_3-y_2}{x_3-x_2}\theta(x-x_2)}_{\text{neuron 2}} + \dots$$



where the ... terms above are identical neurons in both models, with breakpoints $> x_2$ s.t. both models interpolate the data. Then, the slope jumps are $\left|\frac{y_2-y_1}{x_2-x_1}\right| + \left|\frac{y_3-y_2}{x_3-x_2} - \frac{y_2-y_1}{x_2-x_1}\right|$ for model 1 and $\left|\frac{y_3-y_2}{x_3-x_2} - \frac{y_2-y_1}{x_2-x_1}\right|$ for model 2. However, $\tilde{R}$ is

$$\left|\frac{y_2-y_1}{x_2-x_1}\right| + \left|\frac{y_3-y_2}{x_3-x_2} - \frac{y_2-y_1}{x_2-x_1}\right| \quad \text{for model 1 but}$$

$$\left|\frac{y_2-y_1}{x_2-x_1}\right| + \left|\frac{y_3-y_2}{x_3-x_2}\right| \quad \text{for model 2.}$$

We see that model 2 once again has a smaller slope jump but may have larger $\tilde{R}$.

Problems like the two above happen all over. We can fix it by adding a trainable, but not regularized term to the model:

$$z(x;\theta) = ax + b + \sum_{j=1}^{n} W_j^{(2)}\theta(W_j^{(1)}x + b_j^{(1)})$$

This "ax" (a is not regularized), allows for the model to learn a straight line, slope jump-minimizing linear piecewise function (creating models like the two neuron model in 1 or the second model in 2) without needing duplicate break points. It also doesn't change any slope jumps, and so it gives us exactly what we want.

□