Lecture 20: Positive homogeneity & low norm solutions

Reminders.

Clarke differential $\partial f(w) = \left\{ \lim_i \nabla f(w_i) \;:\; \begin{array}{l} w_i \to w \\ \nabla f(w_i) \text{ exists} \\ \lim_i \nabla f(w_i) \text{ exists} \end{array} \right\}$

good: locally Lipschitz $\Rightarrow$ nonempty everywhere

bad: technical issues (a.e. regularity, nonuniqueness of flows)
does not match pytorch

Positive homogeneity: f is L homogeneous $\forall c \geq 0$ $f(cx) = c^L f(x)$.

(e.g., abstraction of L-layer ReLU network without biases)

$$\boxed{\text{key property} \quad \forall s \in \partial f(w), \quad \langle s, w \rangle = L f(w).}$$

$\sigma(c \cdot x) = \max\{0, cx\} = c\,\sigma(x)$

$F(x; cw) = F(x; (cW_2, cb_2, cW_1))$

$= cW_2\,\sigma(cW_1 x) + cb_2$

$= c^2 W_2\,\sigma(W_1 x) + \underbrace{\boxed{cb_2}}$ not pos-hom if $b_2 \neq 0$.

Plan: ① prove & example

② "norm preservation" $\|W_i(t) - W_i(0)\|$ vs. $\|W_k(t) - W_k(0)\|$

③ low norm solutions in optimization

Positive homogeneity & gradients

$$\langle \partial f(w), w\rangle = \{ L f(w)\} \quad ?????$$

ReLU network:

$z_1 := W_1 x$

$z_{i+1} := W_{i+1}\sigma(z_i)$

note $S_i := \mathrm{diag}\left(\boxed{\mathbb{1}[z_i \geq 0]}\right)$, $\sigma(z_i) = \boxed{S_i z_i}$

$$F(x;w) = W_L\sigma(z_{L-1}) = W_L S_{L-1} z_{L-1} = W_L S_{L-1} W_{L-1}\sigma(z_{L-2})$$

$$= \cdots = W_L S_{L-1} W_{L-1} S_{L-2} \cdots S_1 W_1 x.$$

$$\boxed{\frac{d}{dW_j} F(x;w)} = (W_L S_{L-1} W_{L-1}\cdots W_{j+1} S_j)^T (S_{j-1} W_{j-1} \cdots S_1 W_1 x)^T.$$

$W_j \in \mathbb{R}^{d_{j+1}\times d_j}$ (d_{j+1}, d_j)

memory aid
⓪ write out activation matrices
① $\frac{d}{x_j} x_1 x_2 \cdots x_j x_{j+1} \cdots x_n = x_1 \cdots x_{j-1} x_{j+1} \cdots x_n$
② use transposes to match dimensions

$$\langle \partial F(x;w), w\rangle \text{ "=" } \sum_{j=1}^{L} \boxed{\langle \partial_{w_j} F(x;w), W_j\rangle}$$

$$\langle \partial_{w_j} F(x;w), W_j\rangle = \langle (W_L\cdots S_j)^T (S_{j-1}\cdots W_1 x)^T, W_j\rangle$$

$$= tr\left([(W_L \cdots S_j)^T (S_{j-1}\cdots W_1 x)^T]^T W_j \right)$$

$$= tr\left((S_{j-1}\cdots W_1 x)(W_L \cdots S_j) W_j \right)$$

$$= tr\left((W_L \cdots S_j) W_j (S_{j-1} W_{j-1} \cdots W_1 x) \right)$$

$$= tr(F(x;w)) \boxed{= F(x;w).}$$

$$\Rightarrow \langle \partial F(x;w), w\rangle = L\cdot F(x;w)$$

Reminder to future self: instead:
① ReLU network is piecewise affine in the input x, piecewise polynomial in the parameters
② locally Lipschitz ⇒ nondifferentiabilities have Lebesgue measure zero
③ (i) calculate gradient interior of each piece (ii) limits towards boundaries.

Theorem (Euler's homogeneous function theorem). If f is locally Lipschitz $\Rightarrow \forall w$ $\langle \partial f(w), w\rangle = \{L f(w)\}$

Proof. 2-phase proof: ① check when differentiable ② use limits & defn Clarke.

① When differentiable:

$$\langle \nabla f(w), w\rangle = \lim_{h\downarrow 0} \frac{f(w+hw) - f(w)}{h}$$

$$= \lim_{h\downarrow 0} \frac{f((1+h)w) - f(w)}{h}$$

$$= \lim_{h\downarrow 0} \frac{(1+h)^L f(w) - f(w)}{h}$$

$$= \lim_{h\downarrow 0} \frac{(1 + Lh + \binom{L}{2}h^2 \cdots) f(w) - f(w)}{h} = \lim_{h\downarrow 0}\left(Lh f(w) + O(h)\cdots \right)$$

$$= L f(w).$$

$\min_w \sum_i \|w - x_i\|_1$
Clarke book: has drawing of the table

② General case: $s \in \partial f(w) \Rightarrow s = \sum_i \lambda_i s_i$, $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$, $s_i = \lim_{j\to\infty} Df(w_{i,j})$

$$\langle s, w\rangle = \sum_i \lambda_i \langle s_i, w\rangle = \sum_i \lambda_i \langle \lim_j \nabla f(w_{i,j}), w\rangle$$

$$= \sum_i \lambda_i \lim_j \langle \nabla f(w_{i,j}), w_{i,j}\rangle = \sum_i \lambda_i \lim_j L f(w_{i,j})$$

$$= \sum_i \lambda_i L f(w) = L f(w).$$

defn Clarke diff.

//

if $L \in \mathbb{R}_{\geq 0} \setminus \mathbb{Z} \Rightarrow$

$$\lim_{h\downarrow 0} \frac{f((1+h)w) - f(w)}{h}$$

$$= \lim_{h\downarrow 0} \frac{(1+h)^L f(w) - f(w)}{h}$$

$$\text{"="} \lim_{h\downarrow 0} \frac{\exp(hL) f(w) - f(w)}{h}$$

into the trash.

Theorem. Consider GF $\dot{w}_t \in -\partial \hat{R}(w_t) = -\frac{1}{n}\sum_{i=1}^{n} \ell'(y_i F(x_i;w_t)) \partial F(x_i; w_t)$

Suppose this chain rule a.e.

where $F(x;w)$ is 1-homogeneous in each layer

$[\forall c \geq 0$ $F(x; (W_L \cdots, cW_j, W_{j-1}, \cdots, W_1) = c F(x; W_L \cdots W_1)]$

Then $\forall j,k$ $\frac{1}{2}\|W_j(t)\|^2 - \frac{1}{2}\|W_j(0)\|^2 = \frac{1}{2}\|W_k(t)\|^2 - \frac{1}{2}\|W_k(0)\|^2$.

Remark. for other notions of "distance to initialize" might be false.

Proof. $\frac{1}{2}\|W_j(t)\|^2 - \frac{1}{2}\|W_j(0)\|^2 = \int_0^t \frac{d}{ds}\frac{1}{2}\|W_j(s)\|^2 \, ds$

$$= \int_0^t \langle \dot{W}_j(s), W_j(s)\rangle \, ds$$

$$= \int_0^t \langle -\partial_{W_j}\hat{R}(W_j(s)), W_j(s)\rangle \, ds$$

$$= \int_0^t -\frac{1}{n}\sum \ell'(y_i F(x_i;w_s)) \langle \partial_{W_j} F(x_i;W_s), W_j(s)\rangle$$

$$\boxed{= \int_0^t -\frac{1}{n}\sum \ell'(\text{—}) F(x_i;w_s).}$$

no dependence on layer.

$\frac{1}{2}\|W_j(t) - W_j(0)\|^2$

$= \int \langle \dot{W}_j(s), W_j(s) - W_j(0)$

//

[other note to future self: prove clarke decomposes over coordinates]