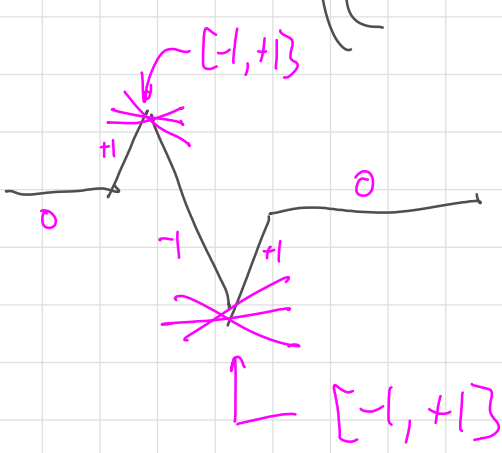
Lec 16: Clarke diff & positive homogeneity

Defn.

$$\partial F(w) := \mathrm{conv}\left(\left\{ \lim_i \nabla F(w_i) : \begin{array}{l} w_i \to w, \\ \nabla f(w_i) \text{ exists} \\ \lim_i \nabla F(w_i) \text{ exists} \end{array} \right\} \right)$$



Basic properties

* F locally Lip $\Rightarrow$ $\partial F \neq \phi$ everywhere
* $F : \mathbb{R}^d \to \mathbb{R}$ convex $\Leftrightarrow$ $\partial F = \partial_c F \neq \phi$
* F differentiable $\Rightarrow$ $\partial F = \{\nabla F\}$

Clarke diff analog of ∇F: "differential inclusion"

$$\dot{w}(t) \in -\partial F(w(t)) \quad \text{a.e. } t \geq 0$$

More serious properties

* (Chain rule.) Under "technical conditions":

$$\text{a.e. } t \geq 0 \qquad \frac{d}{dt} F(w(t)) \begin{cases} \text{if } F = \langle \nabla F(w(t)), \dot{w}(t) \rangle \\ \text{"Clarke flow"} \end{cases}$$

$$= \langle \dot{w}(t), v \rangle \quad \forall v \in \partial F(w(t)).$$

$$\Rightarrow \dot{w}(t) = -\operatorname{argmin}_{v \in \partial F(w(t))} \|v\|$$

$$\Rightarrow \frac{d}{dt} F(w(t)) = -\min_{v \in \partial F(w(t))} \|v\|^2$$

Remark. "technical condition"

① "Whitney Stratifiable" Kakade-Lee, et al?
② "o-minimal definability" Ji-Telgarsky
③ Assume chain rule holds (Lyu-Li)

Hold for standard feedforward networks.

* Under above assumptions,

$$F(w(t)) - F(w(0)) = \int_0^t \frac{d}{ds} F(w(s))\, ds$$

$$= \int_0^t -\min_{v \in \partial F(w(s))} \|v\|^2 \, ds$$

$$\leq -t \cdot \inf_{\substack{s \in [0,t] \\ v \in \partial F(w(s))}} \|v\|^2$$

$$\Rightarrow \inf_{\substack{s \in [0,t] \\ v \in \partial F(w(s))}} \|v\| \leq \sqrt{\frac{F(w(0)) - F(w(t))}{t}}$$

Interpretation: Locally Lip $\Rightarrow$ find $(\sqrt{\varepsilon})$-approx-stationary point in time $\leq t$.

Positive homogeneity

Defn. f is L-positive-homogeneous when $f(cx) = c^L f(x)$ $\forall x$, $\forall c \geq 0$.

Remark Abstraction of multilayer ReLU:

$$f(x; cw) = c W_L \sigma(c W_{L-1} \sigma(c W_{L-2}(\cdots \sigma(cW_1 x)\cdots)$$
$$= (c^L W_L)\sigma(\cdots W_2 \sigma(W_1 x)\cdots)$$
$$= c^L f(x; w).$$

single ReLU is 1-homogeneous
$\sigma(cx) = \max\{c\cdot 0, c\cdot x\} = c\cdot\max\{0, x\} = c\,\sigma(x)$.

Remark

* I don't know of multilayer ReLU results that can't also be proved for general homogeneous
* Unclear how results translate to inhomogeneous case,
* Modern networks are inhomogeneous (e.g., attention).

Positive homogeneity & gradients.

single relu: $\sigma(r) = \max\{0, r\} = r\cdot \mathbb{1}[r \geq 0] = r\cdot\sigma'(r)$

in this equation, behavior at $r=0$ doesn't matter!

Multilayer case: important that we fix the example.

Given x_i define $z_1 := W_1 x$, $z_{i+1} := W_{i+1}\sigma(z_i)$ (pre-activation outputs of layer i.)

$$S_i := \mathrm{diag}(\mathbb{1}[z_i \geq 0])$$

$\Rightarrow \sigma(z_i) = S_i z_i$ (once again behavior at 0 irrelevant)

$$f(x; w) = W_L S_{L-1} W_{L-1} S_{L-2} \cdots S_1 W_1 x$$

this expression unaffected by definition of S when some coordinates of z are 0.

$$\frac{df(x;w)}{dW_i} = (W_L S_{L-1}\cdots S_i)^T (S_{i-1} W_{i-1}\cdots S_1 W_1 x)^T$$

seems to depend on choice of S_i when z_i has some 0 coordinates.

$$\left\langle \frac{df(x;w)}{dW_i}, W_i \right\rangle = \mathrm{tr}\left(\left[\frac{df(x;w)}{dW_i}\right]^T W_i\right)$$
$$= \mathrm{tr}\left(\left[(W_L S_{L-1}\cdots S_i)^T (S_{i-1}\cdots S_1 W_1 x)^T\right]^T W_i\right)$$
$$= \mathrm{tr}\left((S_{i-1}\cdots W_1 x)(W_L S_{L-1}\cdots S_i) W_i\right)$$
$$= \mathrm{tr}\left(W_L S_{L-1}\cdots S_i [W_i] S_{i-1}\cdots S_1 W_1 x\right)$$
$$= f(x; w).$$

consistent

Proposition If g is locally Lipschitz and L-homogeneous, $\forall v \in \partial g(w)$, $\langle v, w\rangle = L\cdot g(w)$.

Proof. First consider $\nabla g(w)$ exists.

$\Rightarrow \forall v \in \partial g(w)$,

$$\langle v, w\rangle = \langle \nabla g(w), w\rangle \quad \text{(chain rule)}$$
$$= \frac{d}{dt} g(w + tw)\Big|_{t=0} \quad \text{(defn derivative)}$$
$$= \lim_{h\to 0} \frac{g(w+hw) - g(w)}{h}$$
$$= \lim_{h\to 0} \frac{(1+h)^L g(w) - g(w)}{h}$$
$$= \lim_{h\to 0} g(w)\left[\frac{(1+h)^L - 1}{h}\right]$$
$$= \lim_{h\to 0} g(w)\left[\frac{1 + Lh + \binom{L}{2}h^2 + h^3 + \cdots - 1}{h}\right]$$
$$= L\cdot g(w).$$

General case: given $v \in \partial g(w)$ where $v = \sum_{i=1}^k \alpha_i v_i$, and $\forall i$ $\exists$ $w_{i,s} \to w$, $\nabla g(w_{i,s})$ exists, $v_i := \lim_s \nabla g(w_{i,s})$ exists

$$\langle v, w\rangle = \left\langle \sum_i \alpha_i v_i, w\right\rangle$$
$$= \sum_i \alpha_i \langle v_i, w\rangle$$
$$= \sum_i \alpha_i \left\langle \lim_s \nabla g(w_{i,s}), \lim_s w_{i,s}\right\rangle$$
$$= \sum_i \alpha_i \lim_s \langle \nabla g(w_{i,s}), w_{i,s}\rangle$$
$$= \sum_i \alpha_i \lim_s L\cdot g(w_{i,s})$$
$$= \sum_i \alpha_i \cdot L\cdot g(w)$$
$$= L\, g(w).$$