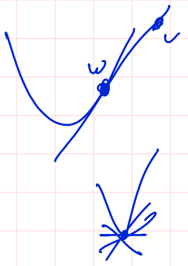
Lecture 17: Clarke & his differential.

Announcement.
* hw2 tomorrow
* notes

why Clarke differential:
* Need a model of a gradient
* Doesn't match pytorch; hard to predict what is used in 10 years.

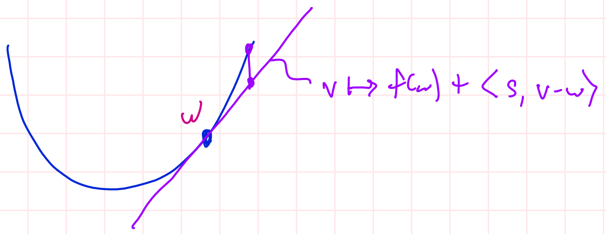
Subgradients

We saw it in optimization:

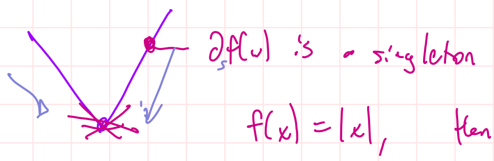


$$\partial_s f(w) = \text{" set of lower bounding slopes at } w\text{"}$$
$$= \{ s \in \mathbb{R}^d : \forall v.\ f(v) \geq f(w) + \langle s, v-w \rangle \}.$$

Examples ①

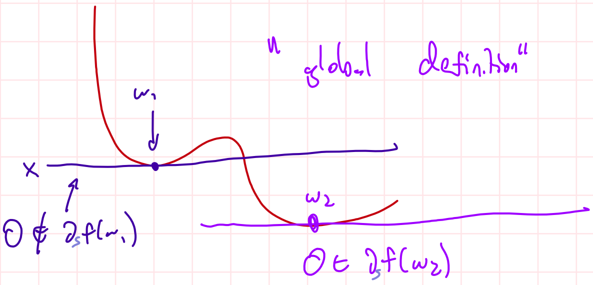


②



$f(x) = |x|$, then $\partial_s f(0) = [-1, +1]$.

③ " global definition"



Properties (for convex $f: \mathbb{R}^d \to \mathbb{R}$).

* $\partial_s f$ is nonempty everywhere, and is closed convex.
* Related to directional derivatives

$$f'(w; v) = \lim_{h \downarrow 0} \frac{f(w+hv) - f(w)}{h} = \sup \{ \langle v, s \rangle : s \in \partial_s f(w) \}$$

* Mean value theorems; see for instance

Hiriart-Urruty & Lemarechal "fundamentals of convex analysis" (many pictures)

Borwein & Lewis "Convex analysis & nonlinear optimization"

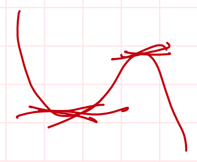
* We used it in our first opt proof.
* Proposition (Jensen's inequality). If $f: \mathbb{R}^d \to \mathbb{R}$ is convex & X is a r.v. then $\mathbb{E} f(X) \geq f(\mathbb{E}X)$.

Proof: $\forall s \in \partial f(\mathbb{E}X)$, $\mathbb{E} f(X) \geq \mathbb{E}[f(\mathbb{E}X) + \langle s, X - \mathbb{E}X \rangle]$

$$= f(\mathbb{E}X) + \langle s, \underbrace{\mathbb{E}X - \mathbb{E}X}_{0} \rangle. \quad //$$

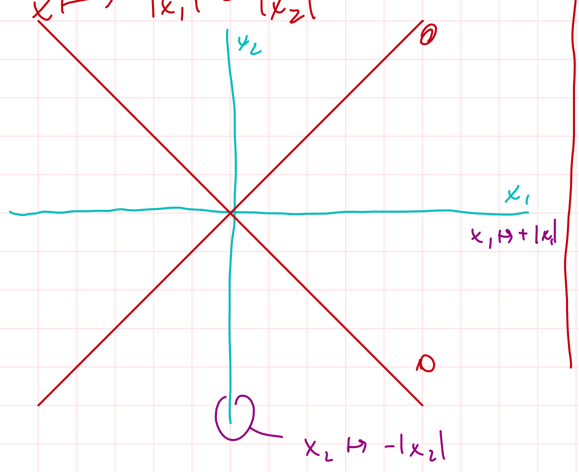
Clarke differential.

First try: "local" sub- supergradient



Seems to work?

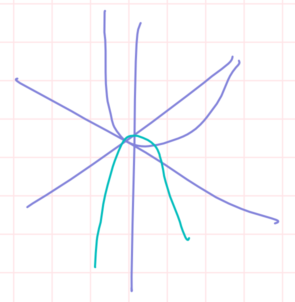
$x \xmapsto{g} |x_1| - |x_2|$



$\mathbb{R}^2$

$\Downarrow$

$x \mapsto x_1^2 - x_2^2$



Want $0 \in \partial g(0)$, fails for "local sub- supergradient"

First definition of Clarke starts from directional derivative

Recall directional derivative: $f'(w;v) = \lim_{h \downarrow 0} \frac{f(w+hv) - f(w)}{h}$

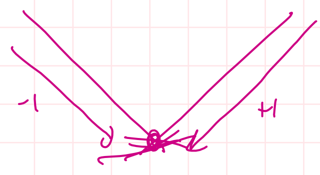
Then $\partial_o f(w) = \{ s \in \mathbb{R}^d : \forall v, \quad f'(w;v) \geq \langle v, s \rangle \}$. ✓

We'll use a related form.

Defn. $\partial f(w) = \underset{\text{convex hull}}{\mathrm{conv}} \left(\left\{ \lim_{i \to \infty} \nabla f(w_i) \;:\; \begin{array}{l} w_i \to w \\ \nabla f(w_i) \text{ exists} \\ \lim \nabla f(w_i) \text{ exists} \end{array} \right\} \right)$

also natural from directional derivative view, with them as boundary points of subgradients.

Example.



$w_i \overset{?}{\to} w$	$\nabla f(w_i) \overset{?}{\to} \lim \nabla f(w_i)$
$1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$	$+1\ +1\ +1\ +1 \to +1$
$-1, -\frac{1}{2}, -\frac{1}{4}, -\frac{1}{8}, \ldots$	$-1\ -1\ -1\ -1 \to -1$
$-1\ +1\ -1\ +1 \not\to$	$-1\ +1\ -1 \not\to$

Convex hull

$\mathrm{conv}(S)$

$= \left\{ \sum_{i=1}^{N} \alpha_i u_i : \begin{array}{l} N \geq 1, \\ u_1, \ldots, u_N \in S \\ \vec{\alpha} \in \Delta \\ (\sum \alpha_i = 1,\ \alpha_i \geq 0) \end{array} \right\}$

Example 2 $\quad g(x) = |x_1| - |x_2|$

claim: $\quad \partial g(0) = \mathrm{conv}(\{ (\pm 1, \pm 1) \})$

$\nabla g(x) = \begin{cases} (\mathrm{sgn}(x_1), \mathrm{sgn}(x_2)) & x_1 \neq 0 \text{ and } x_2 \neq 0 \\ \text{not differentiable} & x_1 = 0 \text{ or } x_2 = 0 \end{cases}$

Remark: Considering any sequence $w_i \to 0$, Clarke differential at 0 seems unstable.

What we want:

1. nonempty everywhere
2. chain rules & gradient flow

---

① Nonempty everywhere

Defn. $f : U \to \mathbb{R}$ (over an open domain U) is locally lipschitz if $\forall x \in U$, $\exists$ open $S \supseteq \{x\}$ s.t. f is Lipschitz over S $\left(\exists M, \ \forall y, z \in S, \quad |f(y) - f(z)| \leq M \|y - z\| \right)$.

picture: $\forall x$, can "zoom in" so that f appears Lipschitz.

Ex. $|x|$ is Lip $\Rightarrow$ locally Lip

x^2 is locally Lipschitz but not Lipschitz

All convex $f : \mathbb{R}^d \to \mathbb{R}$ are locally Lipschitz

$1/x$ over $(0, \infty)$ is locally Lip

$x \sin(1/x)$ over $[0, \infty)$ not locally Lip but uniformly cont.

all standard networks are locally Lipschitz.

Theorem (Rademacher) $\quad f : \mathbb{R}^d \to \mathbb{R}$ is locally Lip $\Rightarrow$ differentiable almost everywhere

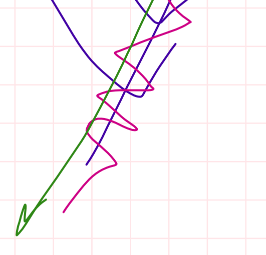
Properties. $f : \mathbb{R}^d \to \mathbb{R}$

1. f locally Lipschitz $\Rightarrow$ ∂f nonempty everywhere
2. f convex $\Rightarrow$ $\partial f = \partial_s f$
3. f differentiable $\Rightarrow$ $\partial f(w) = \{ \nabla f(w) \}$.

Remark. Not what pytorch computes.

$x \mapsto \sigma(x) - \sigma(-x) = x$. (maybe pytorch correct?)

$x \mapsto \sigma(\sigma(x) - \sigma(-x) + 1) - 1$ (maybe not?)



$|x_1| + e^{-x_2}$