Lec 15: nonsmooth, pos-hom, margins

Announcements:
* Schedule until Thanksgiving mostly decided.
* Thursday in Sicbel 1105.
* HW1 graded; check Zinei comments

Plan until Thanksgiving

* Optimization: 4-5 more lectures; non-smooth, positive homogeneity, margins.
* Generalization: cover all main topics before Thanksgiving; thus fair game for homework 4.

After Thanksgiving:

* Hwk4 & project due.
* 3 "extra" lectures.

Next 1.5 lectures:

goal is nonsmooth analog ∇f.

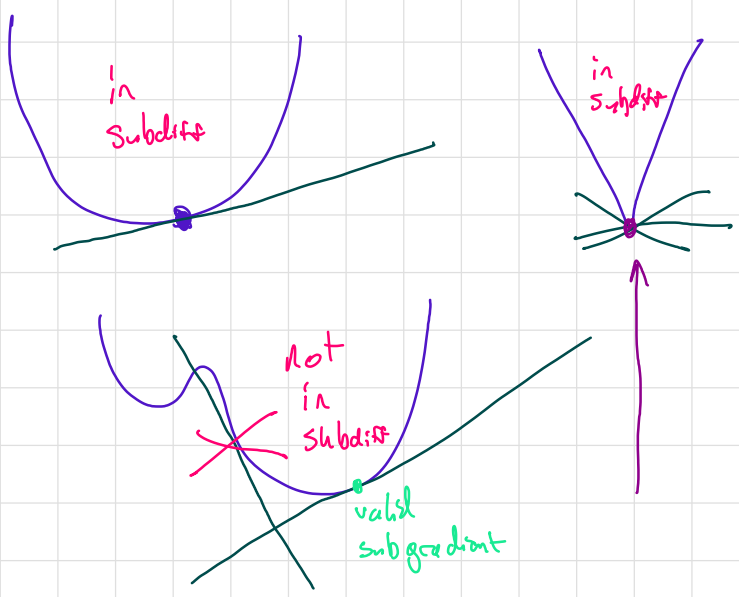
Progression: subgradients → Clarke differential → corresponding flow (positive homogeneity along way)

Subgradients

Define subdifferential / subgradient

$$\partial_c F(w) := \{ s \in \mathbb{R}^d : \forall v \in \mathbb{R}^d,\ F(v) \geq F(w) + \langle s, v-w \rangle \}$$

("set of lower bounding tangents").



Properties: (for convex F.)

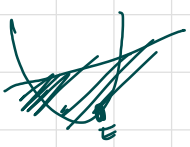
* $\partial_c F$ nonempty everywhere.
  (Convex analysis texts allow $F : \mathbb{R}^d \to \mathbb{R} \cup \{\pm\infty\}$, in which case in general $\partial_c F$ nonempty for $ri(dom(F))$.)
* $\partial_c F$ closed convex set.
* Related to directional derivatives

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

* Most other things you expect from gradients, hold with subgradients of convex functions.

Proposition (Jensen's inequality). Let convex $F : \mathbb{R}^d \to \mathbb{R}$ & r.v. $X \in \mathbb{R}^d$ be given.

Then $F(\mathbb{E}X) \leq \mathbb{E}F(X)$.



Proof. $\exists s \in \partial_c F(\mathbb{E}X)$ (nonempty because convex $F : \mathbb{R}^d \to \mathbb{R}$).

$$\mathbb{E}[F(X)] \geq \mathbb{E}[F(\mathbb{E}X) + \langle s, X - \mathbb{E}X \rangle]$$

defn subgrad applied to integrand.

$$= F(\mathbb{E}X) + \langle s, \mathbb{E}X - \mathbb{E}X \rangle$$

Pretty books on convexity:

④ Fundamentals of convex analysis — Hiriart-Urruty & Lemaréchal — pretty; but long

③ Convex analysis & ("?!") — Borwein & Lewis. — efficient

Clarke differential.

Defn. $F: U \to \mathbb{R}$ ($U \subseteq \mathbb{R}^d$) is *locally Lipschitz* if every open neighborhood is Lipschitz

$$\left(\forall x \in U,\ \exists \text{ open } S \supseteq \{x\}\ \exists M \text{ s.t. } \forall y \in S,\ |f(x) - f(y)| \le M \|x - y\|.\right)$$

E.g. $x \mapsto |x|$ Lipschitz; $1/x$ is locally Lipschitz over $(0,\infty)$, $1/x$ is Lipschitz over $[1,\infty)$.

$x \mapsto x \sin(1/x)$ over $[0,\infty)$: continuous but not locally Lipschitz.

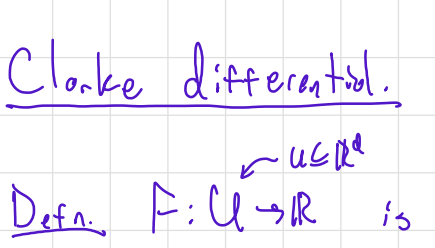
Example: All standard multilayer networks (except threshold) are locally Lipschitz.

Theorem (Rademacher). $F: U \to \mathbb{R}$ locally Lipschitz $\Rightarrow$ differentiable almost everywhere.

Definition. Clarke differential:

$$\partial F(w) = \text{conv}\left(\left\{\lim_{s\to\infty} \nabla F(w_s) : \begin{array}{l} w_s \to w, \\ \nabla f(w_s) \text{ exists}, \\ \lim_{s\to\infty} \nabla F(w_s) \text{ exists.} \end{array}\right\}\right)$$

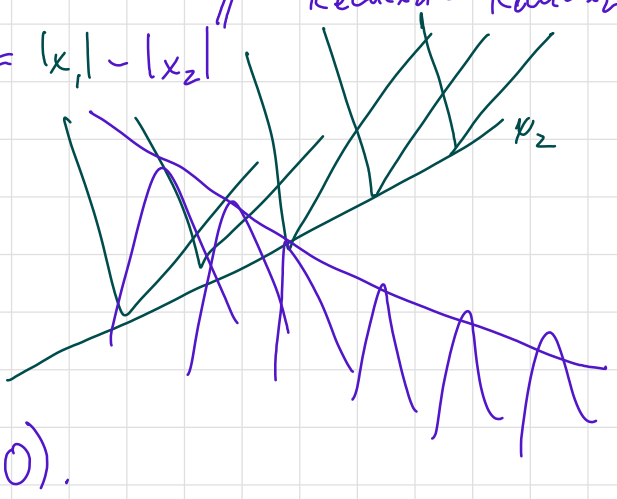
E.g.



another seq with gradients:
$-1, -1, -1, -1, +1, -1, +1, -1, 4$
...

E.g. $F(x) = |x_1| - |x_2|$

$\text{ReLU}(x_1) + \text{ReLU}(-x_1)$
$-\text{ReLU}(x_2) - \text{ReLU}(-x_2)$.



(Claim: $0 \in \partial F(0)$.

Main properties:

* F locally Lipschitz $\Rightarrow \partial F$ exists everywhere.
* $F: \mathbb{R}^d \to \mathbb{R}$ convex $\Rightarrow \partial F = \partial_c f$ everywhere.
* $F: \mathbb{R}^d \to \mathbb{R}$ differentiable everywhere $\Rightarrow \partial f(w) = \{\nabla F(w)\}$ everywhere.

Replace GF: ~~$\dot{w}(t) = \frac{dw}{dt}(t) = -\nabla F(w(t))$~~ $\forall t \ge 0$

Differential inclusion: $\dot{w}(t) = \frac{dw}{dt}(t) \in -\partial F(w(t))$ $\forall$ a.e. $t \ge 0$.

Good news:

* We get a "chain rule"

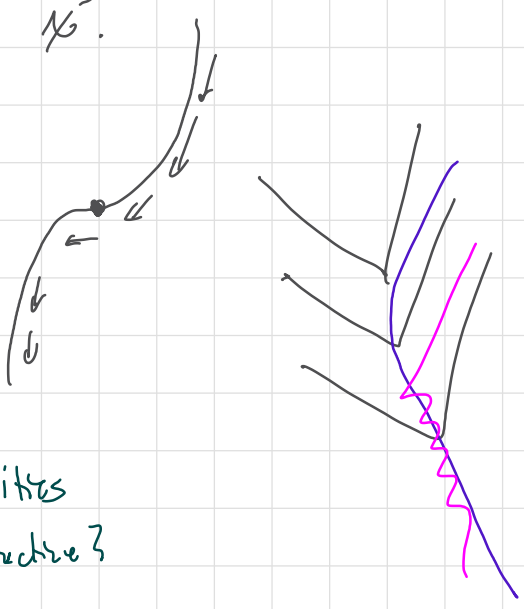
  e.g. $\forall$ a.e. $t \ge 0$ $\quad \frac{d}{dt} F(w(t)) = \langle -v, \dot{w}(t) \rangle$ $\quad \forall v \in \partial F(w(t))$

  Under assumptions (e.g., holds for all standard types of networks)

* Just like regular GF: function values decrease, pass near apx stationary points.

Bad news:

* Does not correspond pytorch etc, not just crazy stuff like $\sigma(\sigma(x)) - \sigma(-x)$, but even x^3.



* Non differentiabilities don't matter practice?