Lecture 23

Last time
- Limiting subdifferential
- Calculus rules

Today
- Clarke subdifferential
- Basic measure theory.

Clarke subdifferential

One awkward issue with the limiting subdifferential is that it doesn't yield convex sets. Recall

Def: For any set $S \subseteq E$

$$\operatorname{conv} S = \left\{ \sum_i \lambda_i x_i \;\middle|\; x_i \in E, \lambda_i > 0, \sum_i \lambda_i = 1 \right\}$$

Caratheodory Theorem
$$= \left\{ \sum_{i=1}^{\dim E+1} \lambda_i x_i \;\middle|\; x_i \in E, \lambda_i > 0, \sum_{i=1}^{\dim E+1} \lambda_i = 1 \right\}$$ ⊣

Proposition: If S is compact, so is $\operatorname{conv} S$. ⊣

Def (Clarke subdifferential): For a locally Lipschitz function $f : E \to \mathbb{R} \cup \{+\infty\}$ define

$$\partial_c f(x) = \operatorname{conv} \partial_L f(x).$$ ⊣

Remark: Clarke introduced this definition and developed theory for it in his PhD thesis in 1973.
This subdifferential is often used in to establish algorithmic convergence.
We will establish an intuitive characterization.

Interlude: Basic facts from Measure Theory

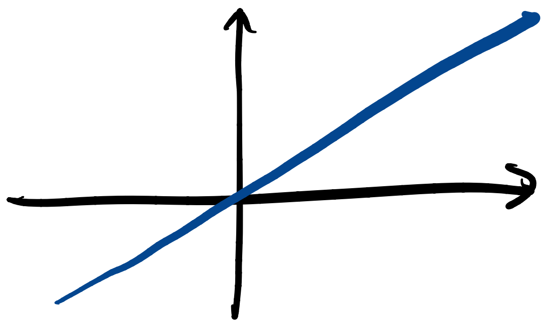
Def: Let $S \subseteq \mathbb{R}^d$. We say that this set has **measure zero** if for all $\varepsilon > 0$ there is a countable sequence of boxes $B_1, B_2, \dots$ with $S \subseteq \bigcup_{i=1}^{\infty} B_i$ s.t. $\sum_i \operatorname{vol} B_i \leq \varepsilon$.

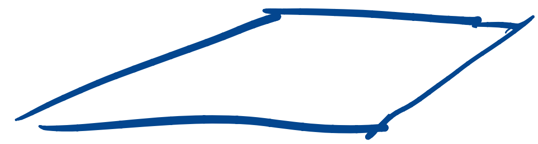
We say that a property holds **almost everywhere (a.e.)** if it holds in the complement of a measure zero set.

Examples

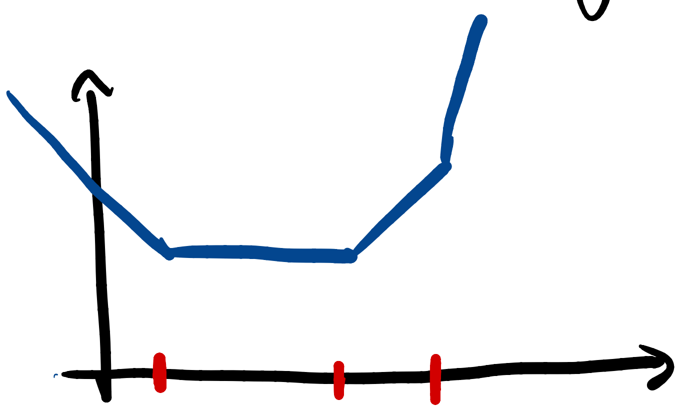
- A countable set has measure zero

- Any affine subspace $S \subseteq E$ with $\dim E > \dim S$ has measure zero.





- Polyhedral functions are differentiable almost everywhere.



This last example is not just a happy coincidence.

Theorem (Radamacher's Theorem): Locally lipschitz functions are differentiable almost everywhere.

We will use one more classical result.

Theorem (Fubini's Theorem): Suppose that $S \subseteq E$ has measure zero, then for almost all x, the set

$$\{ t \in \mathbb{R} \mid x + tz \in S \}$$

has measure zero for all z. $\dashv$

Back to Clarke

Theorem: Let $f : E \to \mathbb{R}$ be Lipschitz and $S \subseteq E$ be the set where f is not differentiable. Then

$$\partial_c f(x) = \underbrace{\operatorname{conv} \{ \lim_{n \to \infty} \nabla f(x_n) \mid x_n \to x,\ x_n \notin S \}}_{\tilde{\partial} f(x)}.$$

Exercise: $\tilde{\partial} f(x)$ is nonempty and compact.

Proof: We start with a claim:

Claim (): $\underset{\swarrow \text{Frechet}}{\partial f(x)} \subseteq \tilde{\partial} f(x)$.

Before proving this claim, let

us see how it implies the result. By the claim $\operatorname{grph} \partial f \subseteq \operatorname{graph} \tilde{\partial} f$. It is easy to show (do it!) that $\operatorname{graph} \tilde{\partial} f$ is closed. By taking the closure of $\operatorname{grph} \partial f$, we obtain that $\operatorname{grph} \partial_L f \subseteq \operatorname{grph} \tilde{\partial} f$. Hence $\partial_L f(x) \subseteq \tilde{\partial} f(x)$ and so

$$\operatorname{conv} \partial_L f(x) \subseteq \tilde{\partial} f(x).$$

For the opposite direction, notice that if f is differentiable at $x_n \Rightarrow \{\nabla f(x_n)\} = \partial f(x)$ and so

$$\operatorname{conv} \partial_L f(x) \supseteq \tilde{\partial} f(x).$$

Proof of Claim $(\square\!\to)$: Suppose seeking contradiction that $\exists\, y \in \partial f(x) \setminus \tilde{\partial} f(x)$. Since $\tilde{\partial} f(x)$ is convex and closed, there is $z \in E$, $\varepsilon > 0$ s.t.

$$\langle y, z \rangle \geq \max_{g \in \tilde{\partial} f(x)} \langle y, z \rangle + 2\varepsilon.$$

So for points $\tilde{x} \in S^c$ close to x

$(*)$ $\langle y, z \rangle \geq \langle \nabla f(\tilde{x}), z \rangle + \varepsilon$. (why?)

By definition $f(x+tz) - f(x) \geq t\langle y, z \rangle + o(t)$.

Thus, for small $t > 0$

$$\frac{1}{t}\left(f(x+tz) - f(x)\right) \geq \langle y, z \rangle - \frac{\varepsilon}{3}$$

By Fubini's there is a point $\bar{x}$ arbitrarily close to x s.t. $\{t \mid \bar{x} + tz \in S\}$ has measure zero. Choose $\bar{x}$ so that

$$\frac{1}{t}\left(f(\bar{x}+tz) - f(\bar{x})\right) \geq \langle y, z \rangle - \frac{2\varepsilon}{3}, \quad (=)$$

which can be done since everything is continuous in x.

Now by the fundamental theorem of calculus

$$f(\bar{x}+tz) - f(\bar{x}) = \int_0^t \langle \nabla f(\bar{x}+\tau z), z \rangle \, d\tau$$

$$\leq t(\langle y, z \rangle + \varepsilon),$$

which contradicts $(=)$. $\square$

This concludes the proof of the theorem. $\square$

Mean Value Theorem

Next we will cover a applications of these subdifferentials and their calculus rules.

Theorem (Mean value 1) Consider a proper closed function $f: \mathbb{E} \to \mathbb{R} \cup \{\infty\}$, and fix two points $x_0, x_1 \in \operatorname{dom} f$. Then, for $\forall \varepsilon > 0$, there exists a subgradient $v \in \partial f(x)$ with $x \in [x_0, x_1] + \varepsilon \mathbb{B}$

$$f(x_1) - f(x_0) \leq \langle v, x_1 - x_0 \rangle + \varepsilon.$$

Proof: Consider the map

$$c(t) = \begin{pmatrix} \overbrace{x_0 + t(x_1 - x_0)}^{c_1(t)} \\ t \\ t \end{pmatrix} \in \begin{pmatrix} \mathbb{E} \\ \mathbb{R} \\ \mathbb{R} \end{pmatrix}$$

and

$$h(y, t_1, t_2) = f(y) - f(x_0) - t_1(f(x_1) - f(x_0)) + \iota_{[0,1]}(t_2)$$

Then, the function

$$\varphi(t) = h(c(t)) = f(c_1(t)) - f(x_0) - t(f(x_1) - f(x_0)) + \iota_{[0,1]}(t)$$

Since this function is proper closed proper and bounded, it admits a minimizer $\hat{t} \in [0,1]$. Applying the fuzzy sum rule we get $\exists\, t_1 \in \mathbb{R}$, $t_2 \in [0,1]$, $x \in E$ s.t.

$$\max\{|t_1 - \hat{t}|, |t_2 - \hat{t}|, \|x - c(\hat{t})\|, |f(x) - f(c(\hat{t}))|\} \leq \varepsilon.$$

and

$$0 \in \langle \partial f(x), x_1 - x_0 \rangle - (f(x_1) - f(x_0)) + N_{[0,1]}(t_2) + [-\varepsilon, \varepsilon]. \qquad (\heartsuit)$$

We consider two cases

- **Case 1: $\hat{t} < 1$.** In that case $(\heartsuit)$ reduces to $\exists\, v \in \partial f(x)$ and $s \in \mathbb{R}_-$ s.t.

$$f(x_1) - f(x_0) \leq \langle v, x_1 - x_0 \rangle + s + \underbrace{q}_{\in [-\varepsilon, \varepsilon]}$$

$$\leq \langle v, x_1 - x_0 \rangle + \varepsilon.$$

▷ Case 2 : $\hat{t} = 1$. Then, noticed that $\varphi(1) = \varphi(0) = 0$ and so 0 is also a minimizer and we can fold back to case 1. $\square$