Subgradients + Subdifferentials of Nonconvex Functions

Let $f:\mathbb{R}^n \to \mathbb{R}$ be continuous (not essential, but simplifies def'ns).

We say y is a REGULAR SUBGRADIENT of f at x $(y \in \hat{\partial} f(x))$ if

$$\liminf_{\substack{z^{(r)} \to 0 \\ z^{(r)} \neq 0}} \frac{f(x+z^{(r)}) - f(x) - y^T z^{(r)}}{\| z^{(r)} \|} \geq 0 .$$

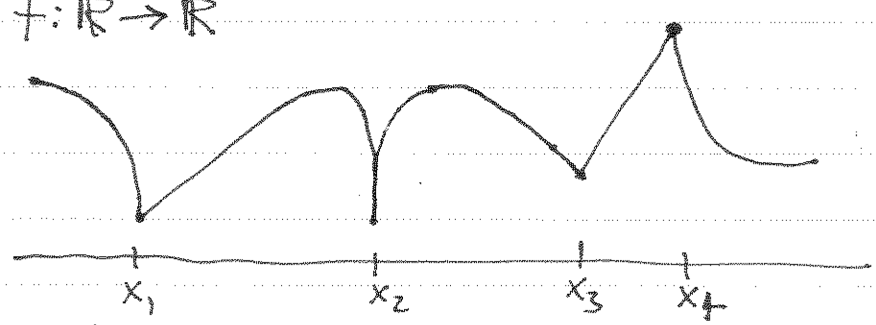
We sometimes write this as

$$f(x+z) \geq f(x) + y^T z + o(\|z\|)$$

Compare this with the convex case, in which $y \in \partial f(x)$ requires $f(x+z) \geq f(x) + y^T z$ MUST HOLD $\forall z$.

Now, a similar inequality must hold for all (SUFFICIENTLY) SMALL z.

e.g. $f:\mathbb{R}\to\mathbb{R}$



The regular subgradients are the slopes of lines that pass through $(x, f(x))$ and lie underneath the graph of f locally, TO FIRST ORDER. Thus

$$\hat{\partial} f(x_1) = (-\infty,], \quad \hat{\partial} f(x_2) = \mathbb{R}, \quad \hat{\partial} f(x_3) = [-1, 1.5], \quad \hat{\partial} f(x_4) = \emptyset .$$

$\hat{\partial} f(x)$ is called the REGULAR SUBDIFFERENTIAL of f at x.

SGNC2

By def'n, $\hat{\partial} f(x)$ is <u>closed</u> and <u>convex</u> but not necessarily compact or nonempty.

In case $n \geq 1$, $\begin{bmatrix} y \\ -1 \end{bmatrix}$ is normal to a hyperplane in $\mathbb{R}^{n+1}$ passing through $(x, f(x))$ and lying underneath the graph of f locally, TO FIRST ORDER.

We say y is a (GENERAL) SUBGRADIENT of f at x $\left(y \in \partial f(x)\right)$ if $\exists \{x^{(r)}\}, \{y^{(r)}\}$ with

$$x^{(r)} \to x$$
$$y^{(r)} \to y$$
$$y^{(r)} \in \hat{\partial} f(x^{(r)}).$$

Clearly $\hat{\partial} f(x) \subseteq \partial f(x)$ (take $x^{(r)} \equiv x$, $y^{(r)} \equiv y \in \hat{\partial} f(x)$)

In our example, for what x is $\partial f(x) \neq \hat{\partial} f(x)$?

Answer: only x_4, with $\partial f(x_4) = \{1.5, -1.5\}$ NOT A CONVEX SET.

We say y is a HORIZON SUBGRADIENT of f at x $\left(y \in \partial^\infty f(x)\right)$ if $\exists \{x^{(r)}\}, \{y^{(r)}\} \in \mathbb{R}^n$, $\{t_r\} \in \mathbb{R}_+$ with

$$x^{(r)} \to x$$
$$t_r y^{(r)} \to y, \quad t_r \to 0 \ (\text{i.e., } t_r \downarrow 0)$$
$$y^{(r)} \in \hat{\partial} f(x^{(r)}).$$

If $\hat{\partial} f(x) \neq \emptyset$, then $\underline{0} \in \partial^\infty f(x)$ (take $x^{(r)} \equiv x$, $y^{(r)} \in \hat{\partial} f(x)$, $t_r \equiv 0$)

In our example, for what x is $\partial^\infty f(x) \neq \{0\}$?

Answer: x_1, x_2 : $\partial^\infty f(x_1) = (-\infty, 0)$, $\partial^\infty f(x_2) = \mathbb{R}$.

SGNC 3

We call $\partial f(x)$ the subdifferential of f at x and $\partial^\infty f(x)$ the horizon subdifferential of f at x.

If f is convex or f is C^1 (continuously differentiable) at x, then

$$\partial f(x) = \hat{\partial} f(x), \quad \partial^\infty f(x) = \{0\}.$$

Note: If $f(x) = -|x|$, then $\hat{\partial} f(0) = \phi$, $\partial f(0) = \{-1, 1\}$, and $\partial^\infty f(0) = \{0\}$.

__Simplest Nontrivial Example__

(Lewis, "Nonsmooth Analysis of Eigenvalues")

Let $\varphi_k(x) = k^{th}$ largest element of $\{x_1, \ldots, x_n\}$.
$= x_{[k]}$ in BV notation.

Clearly φ_k is convex iff $k = 1$.

__Thm__
$$\hat{\partial} \varphi_k(x) = \begin{cases} \text{conv}\{e^i : x_i = \varphi_k(x)\} & \text{if } k = 1 \text{ or} \\ \{k > 1 \text{ and } \varphi_{k-1}(x) > \varphi_k(x) \\ \phi \text{ otherwise} \end{cases}$$

Here $e^i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i$

e.g. $x = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 2 \\ 3 \end{bmatrix} \in \mathbb{R}^5$

$$\hat{\partial} \varphi_1(x) = \left\{ \begin{bmatrix} 0 \\ \tau \\ 0 \\ 0 \\ 1-\tau \end{bmatrix} : \tau \in [0, 1] \right\}$$

$$\hat{\partial} \varphi_2(x) = \phi \qquad \text{etc.}$$

SGNC4

Pf Let $J = \{i : x_i = \varphi_k(x)\}$

If $k=1$, then φ_k is convex so

$$\hat{\partial}\varphi_1(x) = \partial\varphi_1(x) = \{y : \varphi_1(x+z) \geq \varphi_1(x) + y^T z \quad \forall z \in R^n\}$$
$$= \mathrm{conv}\{e^i : i \in J\}.$$

If $k>1$ and $\varphi_{k-1}(x) > \varphi_k(x)$ then (sufficiently) close to x, φ_k is equivalent to

$$\omega \longmapsto \max_{i \in J}(\omega_i)$$

This is convex with subdifferential $\mathrm{conv}\{e^i : i \in J\}$ so this set is $\hat{\partial}\varphi_k(x)$.

On the other hand, if $\varphi_{k-1}(x) = \varphi_k(x)$, we have $|J| \geq 2$.

Suppose $\exists y \in \hat{\partial}\varphi_k(x)$ so

$$\varphi_k(x+z) \geq \varphi_k(x) + y^T z + o(z) \quad \text{(for any sequence } z \to 0)$$

For any index $i \in J$, all suff small $\delta > 0$, we have

$$\varphi_k(x + \underbrace{\delta e^i}_{z}) = \varphi_k(x)$$

since the perturbation δe^i changes only one of the two or more entries equal to $\varphi_k(x)$, so $y_i \leq 0$. Also

$$\varphi_k\Big(x \underbrace{- \delta \sum_{i \in J} e^i}_{z}\Big) = \varphi_k(x) - \delta$$

since the perturbation changes ALL the entries equal to $\varphi_k(x)$

so $y^T z = -\delta \sum_{i \in J} y_i \leq \delta$, i.e. $\sum_{i \in J} y_i \geq 1$: CONTRADICTION.

$\therefore \hat{\partial}\varphi_k(x) = \emptyset$

SGNC5

Not hard to see that $\partial^\infty \varphi_k(x) = \{0\} \ \forall x$.

<u>Thm.</u> $\partial\varphi_k(x) = \{y : y \in \operatorname{conv}\{e^i : x_i = \varphi_k(x)\}$

<u>and</u> $(\#\, y_i \text{ that are } > 0) \le \underbrace{(\#\, x_i \text{ that are } \ge \varphi_k(x)) - k + 1}_{(*)}\}$

e.g. $x = \begin{bmatrix}1\\3\\2\\2\\3\end{bmatrix}$ $\quad \partial\varphi_1(x) = \left\{\begin{bmatrix}0\\\tau\\0\\0\\1-\tau\end{bmatrix} : \tau \in [0,1]\right\}$ as $(*) = 2-1+1 = 2$

(which we already knew)

$$\partial\varphi_2(x) = \left\{\begin{bmatrix}0\\1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}\right\} \text{ as } (*) = 2-2+1 = 1$$

NOT CONVEX.

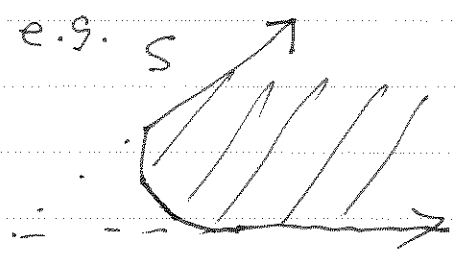
<u>Pf</u> Long. See Lewis, NSAE.

<u>Def</u>. The $\left\{\begin{matrix}\text{recession cone}\\ \text{horizon cone}\end{matrix}\right\}$ of a nonempty, closed, convex set S, denoted S^∞, is

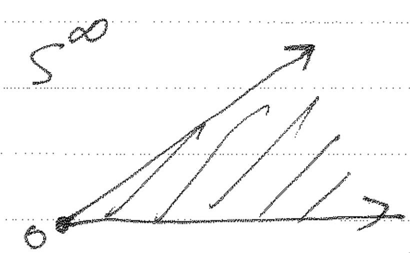
$$\{v : w + tv \in S \ \forall t \in \mathbb{R}_+\}$$

where w is <u>any</u> given element of S.

e.g.



(doesn't matter whether $0 \in S$)



(always $0 \in S^\infty$)

e.g. $\{[-1,\infty)\}^\infty = \{[1,\infty)\}^\infty = \mathbb{R}_+$.

REGULARITY.

f is $\left\{\begin{matrix}\text{Clarke} \\ \text{subdifferentially}\end{matrix}\right\}$ REGULAR at x if

1. $\partial f(x) = \hat{\partial} f(x) \neq \emptyset$ (not standard, assume for simplicity)
2. $\partial^\infty f(x) = \left(\hat{\partial} f(x)\right)^\infty$ = horizon cone of $\hat{\partial} f(x)$.

For our example, for what x is f not regular at x?
Answer only x_4.

Facts f is convex. $\Rightarrow$ f is regular at all x
f is C^1 at x $\Rightarrow$ f is regular at x.

(Differentiable is not enough, e.g. $f(x) = \begin{cases} x^2 \sin \frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$)

Thus regularity generalizes smoothness (C^1) AND convexity.

e.g. φ_k (k^{th} largest element) is regular at x iff $k = 1$ or $k > 1$ and $\varphi_{k-1}(x) > \varphi_k(x)$ (corollary of previous thm on p. SGN3.)

SGNC7

Why do we care about regularity?

<u>THM</u> (Chain Rule – simplest version)

Let $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^{n}$.

Let $f : \mathbb{R}^n \to \mathbb{R}$, and define $h : \mathbb{R}^m \to \mathbb{R}$ by

$$h(\xi) = f(A\xi + b)$$

Suppose

(1) f is regular at $A\bar{\xi} + b$ for some $\bar{\xi} \in \mathbb{R}^m$

(2) $A^T y = 0$ and $y \in \partial^\infty f(A\bar{\xi} + b) \Rightarrow y = 0$

i.e. $\mathcal{N}(A^T) \cap \partial^\infty f(A\bar{\xi} + b) = \{0\}$.

Then h is regular at $\bar{\xi}$, with

$$\partial h(\bar{\xi}) = A^T \partial f(A\bar{\xi} + b)$$

$$\partial^\infty h(\bar{\xi}) = A^T \partial^\infty f(A\bar{\xi} + b).$$

<u>Pf</u> Rockafellar + Wets, Springer 1998, Chapter 10.

Extremely useful property.

<u>Def</u> f is Lipschitz (with const L) on a set S
if $\|f(x)-f(y)\| \le L\|x-y\| \quad \forall x,y \in S$.

If f is Lipschitz on a <u>nbhd</u> of $z \in \mathbb{R}^n$, we say
f is <u>locally Lipschitz</u> at z.

In our example, at which x is f not locally Lipschitz?
<u>Answer</u>. x_1, x_2.

<u>Fact</u> If f is locally Lipschitz at x, then $\partial^\infty f(x) = \{0\}$.

Suppose f is locally Lipschitz at x. We say that
y is a Clarke subgradient of f at x (written
$y \in \partial^c f(x) = \{$generalized gradient$\}$ of f at x if
it is a convex combination of subgradients of f at x,
i.e.

$$\partial^c f(x) = \operatorname{conv}(\partial f(x)).$$

e.g. $f(x) = -|x|$: $\partial^c f(x) = \operatorname{conv}(\{-1,+1\}) = [-1,1]$

<u>Fact</u> $\partial^c f(x) = \operatorname{conv} G(x)$

where $G(x) = \{ g : \exists x^{(k)} \to x$, f is differentiable at $x^{(k)}$
with $\nabla f(x^{(k)}) \to g \}$.

<u>Note</u> If f is locally Lipschitz and regular at x,
then $\partial^c f(x) = \partial f(x) = \hat{\partial} f(x)$ and $\partial^\infty f(x) = \{0\}$.

i.e. all 3 kinds of subgradients
(Clarke, "general", regular) are the SAME.

SGNC9

Optimality Conditions

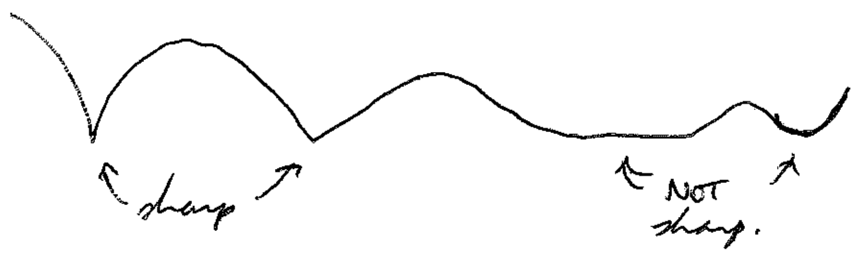
$0 \in \hat{\partial} f(\bar{x})$ is a NEC condition for $\bar{x}$ to locally minimize f.

Pf: HW (easy).

Sufficient conditions are more complicated. Here is a very strong one.

Def $\bar{x}$ is a sharp local minimizer of f if $\exists\, \tau > 0$ s.t.

$$(*) \quad f(\bar{x}+z) - f(\bar{x}) \geq \tau \|z\| \quad \forall z \text{ with } \|z\| \text{ suff. small.}$$



Thm. $\bar{x}$ is a sharp local minimizer of f iff $0 \in \underline{\text{int}}\, \hat{\partial} f(\bar{x})$.

Pf. $(\Rightarrow)$. We have, $\forall w$ with $\|w\| \leq 1$,

$$\frac{f(\bar{x}+z) - f(\bar{x}) - \tau w^T z}{\|z\|} \underset{\text{by } (*)}{\geq} \frac{\tau \|z\| - \tau w^T z}{\|z\|} \geq 0$$

so taking lim inf, $\tau w \in \hat{\partial} f(\bar{x})$, so $\tau B \in \hat{\partial} f(\bar{x})$ where $B = \{w : \|w\| \leq 1\}$ (unit ball).

$\therefore 0 \in \text{int}\, \hat{\partial} f(\bar{x})$.

$(\Leftarrow)$. Suppose $0 \in \text{int}\, \hat{\partial} f(\bar{x})$, so $\exists\, \sigma > 0$ with $\sigma B \in \hat{\partial} f(\bar{x})$, so

$$\liminf_{z^{(k)} \to 0} \frac{f(\bar{x}+z^{(k)}) - f(\bar{x}) - \sigma w^T z^{(k)}}{\|z^{(k)}\|} \geq 0 \quad \forall w \in B.$$

incl. w with $\|w\| = 1$.

In particular, this is true for the sequence $z^{(r)} = \delta_r w$
$\delta_r \in \mathbb{R}$, $\delta_r \downarrow 0$, $\|w\| = 1$.

So

$$\liminf_{\delta_r \downarrow 0} \frac{f(\bar{x} + \delta_r w) - f(\bar{x})}{\delta_r} \geq \sigma$$

Let $\tau < \sigma$. Then, for suff. small δ_r

$$\frac{f(\bar{x} + \delta_r w) - f(\bar{x})}{\delta_r} \geq \tau .$$

Since this is true $\forall w$ with $\|w\| = 1$, we have $(*)$.