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

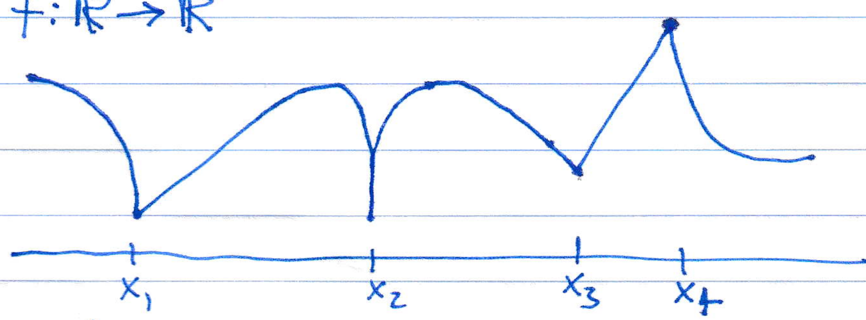
$$f(x+z) \geq f(x) + y^T z + o(\|z\|)$$

Compare this with the convex case, in which $y \in \partial f(x)$ requires

$$f(x+z) \geq f(x) + y^T z \quad \text{MUST HOLD } \forall z.$$

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

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

$\hat{\partial} f(x)$ 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 ($y \in \partial f(x)$) 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 $\partial f(x_4) = \{1.5, -1\}$ NOT A CONVEX SET.

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

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

and if $\hat{\partial} f(x)$ is bounded, $\partial^\infty f(x) = \{0\}$.

If $\hat{\partial} f(x) \neq \emptyset$, then $0 \in \partial^\infty f(x)$ (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 : $\partial^\infty f(x_1) = (-\infty, 0)$, $\partial^\infty f(x_2) = \mathbb{R}$.

SGNC3

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 $\hat{\partial} f(0) = \emptyset$, $\partial f(0) = \{-1, 1\}$, and $\partial^\infty f(0) = \{0\}$.
NOT $[-1, 1]$

<u>Simplest Nontrivial Example</u>
(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 $k=1$.

<u>Thm</u> $$\hat{\partial}\varphi_k(x) = \begin{cases} \operatorname{conv}\{e^i : x_i = \varphi_k(x)\} & \text{if } k=1 \text{ or } \{k>1 \text{ and } \varphi_{k-1}(x) > \varphi_k(x)\} \\ \emptyset & \text{otherwise} \end{cases}$$

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

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

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

$\hat{\partial}\varphi_2(x) = \emptyset$ etc. $\hat{\partial}\varphi_3(x) = \left\{ \begin{bmatrix} 0 \\ 0 \\ \tau \\ 1-\tau \\ 0 \end{bmatrix} : \tau \in [0,1] \right\}$

SGNC4

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

If $k=1$, then φ_k is convex so

$$\hat{\partial}\varphi_1(x) = \partial\varphi_1(x) = \{y : \varphi_1(x+z) \geq \varphi_1(x) + y^T z \quad \forall z \in R^n\}$$

$$= \text{conv}\{e^i : i \in J\}.$$

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

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

This is convex with subdifferential $\text{conv}\{e^i : i \in J\}$ 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, $y_i \leq 0 \ \forall i \in J$ ← as $y^T z = \delta y_i$

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

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

Since $\hat{\partial} \varphi_k(x)$ is bounded, clearly $\partial^\infty \varphi_k(x) = \{0\}\ \forall x$, by def'n.

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

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

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.

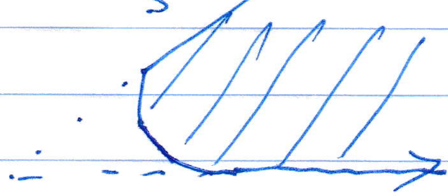
Pf Long. See Lewis, NSAE.

Def. 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 any given element of S.

e.g. S



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

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)

and

2. $\partial^\infty f(x) = (\hat{\partial} f(x))^\infty$ = 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

CHAIN RULE

R+W Thm 10.6 Simplified.

Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be smooth, i.e. C^1

with Jacobian
$$\underset{(m\times n)}{F'} = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \vdots & & \\ \frac{\partial F_m}{\partial x_1} & \cdots & \frac{\partial F_m}{\partial x_n} \end{bmatrix}$$
(called ∇F in R+W).

and let $g : \mathbb{R}^m \to \mathbb{R}$ be continuous.

Suppose further that, at $\bar{x} \in \mathbb{R}^n$,

(1) g is regular at $F(\bar{x})$

(2) $\mathcal{N}(F'(\bar{x})^T) \cap \partial^\infty g(F(\bar{x})) = \{0\}$.

Define $f = g \circ F$, so $f(x) = g(F(x))$ $f : \mathbb{R}^n \to \mathbb{R}$.

Then f is regular at $\bar{x}$, with

$$\partial f(\bar{x}) = \underbrace{\underbrace{(F'(\bar{x}))^T}_{n\times m} \underbrace{\partial g(F(\bar{x}))}_{\subseteq \mathbb{R}^m}}_{\subseteq \mathbb{R}^n} \qquad (\dagger)$$

and

$$\partial^\infty f(\bar{x}) = (F'(\bar{x}))^T \partial^\infty g(F(\bar{x}).$$

See Rockafellar + Wets, "Variational Analysis", Springer 1999, Ch. 10.

Very useful property!

SGNC8

__Def__ f is Lipschitz (with const L) on a set S if $\|f(x)-f(y)\| \leq L\|x-y\| \quad \forall x,y \in S$.

If f is Lipschitz on a __nbhd__ of $z \in \mathbb{R}^n$, we say f is __locally Lipschitz__ at z.

In our example, at which x is f not locally Lipschitz?

__Answer__. x_1, x_2.

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

__Fact__ $\partial^C f(x) = \operatorname{conv} G(x)$

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

__Note__ 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.

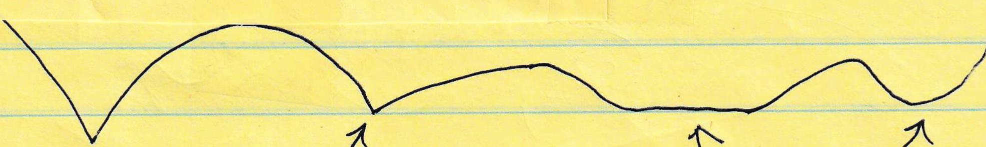
$0 \in \hat{\partial} f(x)$ is a necessary condition for x to locally minimize f.

Pf: immediate from def'n of $\hat{\partial} f$ (HW).

Sufficient conditions are more complicated. Here is one with strong assumptions + a strong consequence.

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

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



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

Pf ($\Rightarrow$) We have, $\forall w \in \mathbb{R}^n$ with $\|w\|_2 \leq 1$,

$$\frac{f(x+z) - f(x) - \tau w^T z}{\|z\|_2} \underset{\text{by } (*)}{\geq} \frac{\tau \|z\|_2 - \tau w^T z}{\|z\|_2} \underset{\text{uses } \|w\|_2 \leq 1.}{\geq} 0$$

so taking lim inf over $z = z^{(k)}$,

$$\tau w \in \hat{\partial} f(x), \text{ so } \tau B \subseteq \hat{\partial} f(x)$$

where $B = \{w : \|w\|_2 \leq 1\}$ (unit ball)

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

($\Leftarrow$) next page.

SGNC10

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

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

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

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

Let $\tau < \sigma$ ($\tau > 0$). Then, for suff. small δ_n

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

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