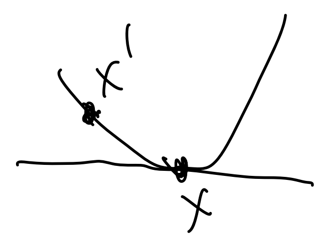
Clarke Differential

Subdifferential and Subgradient

Definition: Given $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the subdifferential set is defined as

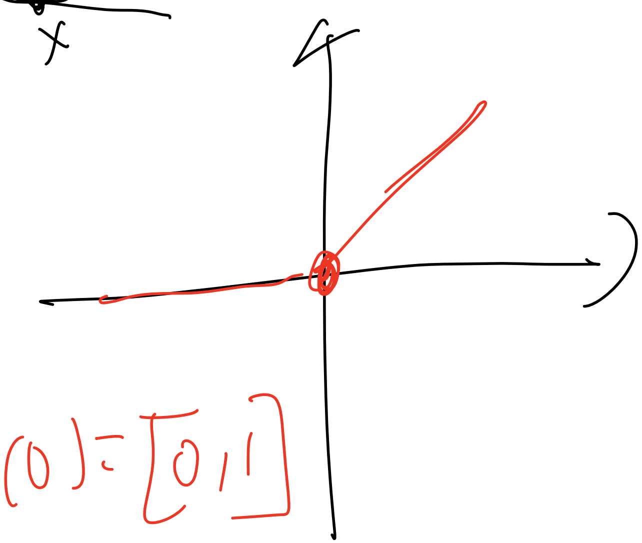
linear function

$\partial_s f(x) \triangleq \{s \in \mathbb{R}^d : \underline{\forall x' \in \mathbb{R}^d}, f(x') \geq f(x) + \underline{s^\top(x' - x)}\}$. The elements in the subdifferential set are subgradients.



$$x_{t+1} = x_t - \eta g_t$$

$$g_t \in \partial_s f(x)$$



$$\partial_s f(0) = [0, 1]$$

Subdifferential and Subgradient

Definition: Given $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the subdifferential set is defined as $\partial_s f(x) \triangleq \{s \in \mathbb{R}^d : \forall x' \in \mathbb{R}^d, f(x') \geq f(x) + s^\top(x' - x)\}$. The elements in the subdifferential set are subgradients.

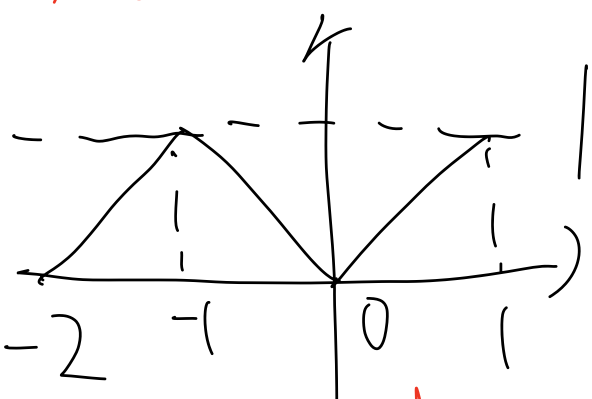
- If f is convex $\Rightarrow$ $\partial_s f(x)$ exists everywhere
- If f is convex and differentiable $\partial_s f(x) = \{\nabla f\}$ & smooth $\Downarrow$ $O\left(\frac{1}{T}\right)$

$O\left(\frac{1}{\sqrt{T}}\right)$ rate

Subdifferential is not enough

Definition: Given $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the subdifferential set is defined as $\partial_s f(x) \triangleq \{s \in \mathbb{R}^d : \forall x' \in \mathbb{R}^d, f(x') \geq f(x) + s^\top (x' - x)\}$. The elements in the subdifferential set are subgradients.

Problem: NN is not convex



$x = -1$

we need s s.t. $\forall x'$

$f(x') \geq 1 + s \cdot (x' - (-1))$

- choose $x' = -2$

$0 \geq 1 + s \cdot (-1) \Rightarrow s \geq 1$

- choose $x' = 1$

$1 \geq 1 + s \cdot (2) \Rightarrow s \leq 0$

$\partial_s f(-1)$: empty set

$\Rightarrow$ well-defined

Clarke Differential

$x_{t+1} = x_t - \eta_t g_t$

$g_t \in \partial f(x)$

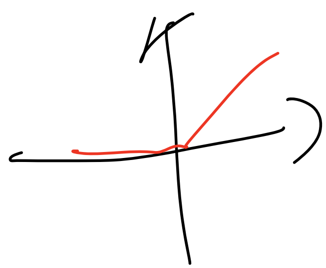
Definition: Given $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the Clark differential is defined as

$$\partial f(x) \triangleq \text{conv}\left(\{s \in \mathbb{R}^d : \exists \{x_i\}_{i=1}^{\infty} \to x, \{\nabla f(x_i)\}_{i=1}^{\infty} \to s\}\right).$$

The elements in the subdifferential set are subgradients.

$\text{conv}(S) = \{v : v = \sum_{i=1}^{n} \lambda_i u_i,\ u_i \in S,\ \lambda_i \geq 0,\ \sum_{i=1}^{n} \lambda_i = 1\}$

example: ReLU



$\{x_i\},\ -1,\ -\frac{1}{2},\ \ldots \to 0$

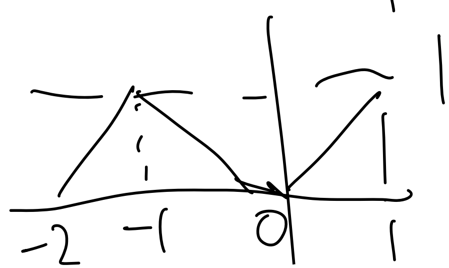
$\nabla f(x_i) = 0$

$\{x_i\},\ 1,\ \frac{1}{2},\ \ldots \to 0$

$\nabla f(x_i) = 1$

$\text{conv}(\{0, 1\}) = [0, 1]$

example:



$\{x_i\}: -2,\ -1.5,\ \ldots \to -1,\ \nabla f(x_i) = 1 \Rightarrow \text{conv}(\{1, -1\})$

$\{x_i\}: 0,\ -\frac{1}{2},\ \ldots \to -1,\ \nabla f(x_i) = -1 \quad = [-1, 1]$

When does Clarke differential exists

$x: \exists S, \forall x', |f(x)-f(x')| \le L\|x-x'\|$

Definition (Locally Lipschitz): $f : \mathbb{R}^d \to \mathbb{R}$ is locally Lipchitz if $\forall x \in \mathbb{R}^d$, there exists a neighborhood S of x, such that f is ~~Lipchitz in S~~.

- If f is locally Lipshitz $\to \partial f$ exits everywhere
- If f is convex $\Rightarrow \partial f = \partial_s f$
- If f is differentiable $\Rightarrow \partial f = \{\nabla f\}$
- If f is

Positive Homogeneity

$\sigma(z) = \max\{z, 0\}$

Definition: $f : \mathbb{R}^d \to \mathbb{R}$ is positive homogeneous of degree L if $f(\alpha x) = \alpha^L f(x)$ for any $\alpha \geq 0$.

① ReLU: $\sigma(\alpha z) = \alpha \cdot \sigma(z)$

② monomials of degree L: $\prod_{i=1}^{d} x_i^{p_i}$, $\sum_{i=1}^{d} p_i = L$

$$\prod_{i=1}^{d} (\alpha x_i)^{p_i} = \alpha^{\sum p_i} \cdot \prod x_i$$

$$= \alpha^L \prod_{i=1}^{d} x_i^{p_i}$$

③ Norm: $\|\alpha x\| = \alpha \cdot \|x\|$

Positive Homogeneity

(4) Multi-layer ReLU

$f(x, W_1, \ldots, W_{H+1}) = W_{H+1}\,\sigma(W_H \cdots \sigma(W_1 x) \cdots)$

— for one-layer

$f(x, W_1, \ldots, \alpha W_h, \ldots, W_{H+1}) = \alpha \cdot W_{H+1}\,\sigma(W_H \cdots \sigma(W_1 x) \cdots)$

— for all layers

$f(x, \alpha W_1, \ldots, \alpha W_{H+1}) = \alpha^{H+1} f(x, W_1, \ldots, W_{H+1})$

$\Rightarrow$ $(H+1)$ - homogeneous

Positive Homogeneity

say $W_h \in \mathbb{R}^{m \times m}$

Fact: $\forall h = 1, \ldots, H+1$

$$\left\langle W_h, \frac{\partial f(x, W_1, \ldots, W_{H+1})}{\partial W_h} \right\rangle = \underbrace{f(x, W_1, \ldots, W_{H+1})}_{\text{independent of } h}$$

Pf: $A_h = \mathrm{diag}\left(\sigma'(W_h \sigma(\cdots \sigma(W_1 x) \cdots))\right) \in \mathbb{R}^{m \times m}$

$\sigma' = 0$ or $1 \Rightarrow$ pattern whether activation is on or not

$$f(x, W_1, \ldots, W_{H+1}) = W_{H+1} A_H W_H \cdots A_1 W_1 x$$

$$\frac{\partial f}{\partial W_h} = (W_{H+1} A_H \cdots W_{h+1} A_h)^T (A_{h-1} W_{h-1} \cdots W_1 x)$$

$\Rightarrow$ verify

Positive Homogeneity and Clark Differential

Lemma: Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is Locally Lipschitz and L -positively homogeneous. For any $x \in \mathbb{R}^d$ and $s \in \partial f(x)$, we have $\langle s, x \rangle = Lf(x)$.

if choose $x = W_n$

$s : \frac{\partial f}{\partial W_n}$

$\langle s, x \rangle = f(x)$

Norm Preservation

$f(X, W_1, W_2, W_3) = W_3\, \sigma(W_2\, \sigma(W_1 X))$

quadratic loss

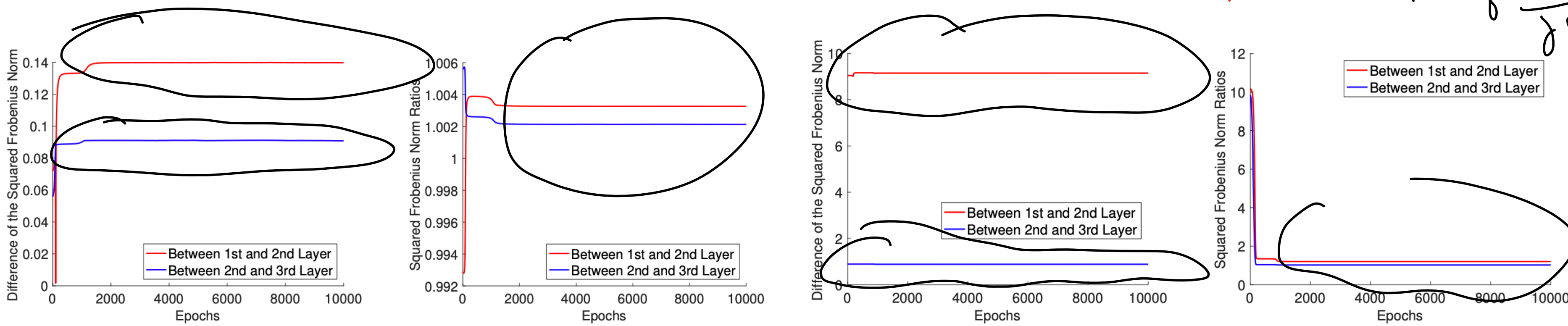
$W_3 \times 10$

$W_2 / 10$

$W_2(t+1) = W_2(t) - \eta \frac{\partial L}{\partial W_2}$

$W_3(t+1) = W_3(t) - \eta \frac{\partial L}{\partial W_3}$

constant



(a) Balanced initialization, squared norm differences.

(b) Balanced initialization, squared norm ratios.

(c) Unbalanced Initialization, squared norm differences.

(d) Unbalanced initialization, squared norm ratios.

(a) — $\|W_1\|_F^2 - \|W_2\|_F^2$

(b) — $\dfrac{\|W_1\|_F^2}{\|W_2\|_F^2}$

$\|W_1\|_F^2 - \|W_1(0)\|_F^2 = \|W_2\|_F^2 - \|W_2(0)\|_F^2$

if init small

$\Rightarrow$ balanced layers

$W_1(0)$, $W_2(0)$ small

$\Rightarrow \|W_1(t)\|_F^2 \approx \|W_2(t)\|_F^2$

Gradient flow and gradient inclusion

Discrete-time dynamics can be complex. Let's use continuous-time dynamics to simplify:

Gradient flow: $x_{t+1} = x_t - \eta \nabla f(x_t) \Rightarrow \dfrac{x(t)}{dt} = -\nabla f(x(t))$

Gradient inclusion: $\dfrac{dx(t)}{dt} \in \partial f(x(t))$

$$\frac{x_{t+1} - x_t}{\eta} = -\nabla f(x)$$

let $\eta \to 0$

Norm preservation by gradient inclusion

Theorem (Du, Hu, Lee '18) Suppose $\alpha > 0$, $f(x; (W_{H+1}, \ldots, \alpha W_i, \ldots, W_1)) = \alpha f(x, (W_{H+1}, \ldots, W_1))$, I.e., predictions are 1-homogeneous in each layer. Then for every pair of layers $(i, j) \in [H+1] \times [H+1]$, the gradient inclusion maintains: for all $t \geq 0$,

$$\frac{1}{2}\|W_i(t)\|_F^2 - \frac{1}{2}\|W_i(0)\|_F^2 = \frac{1}{2}\|W_j(t)\|_F^2 - \frac{1}{2}\|W_j(0)\|_F^2.$$

$\Rightarrow$ $\|W_i(t)\|_F^2 \approx \|W_j(t)\|_F^2$ if init is small

Pf: $\int \frac{d\|W_i(\tau)\|_F^2}{d\tau} d\tau$, $\frac{d\|W_i(\tau)\|_F^2}{d\tau} = 2\left\langle W_i(\tau), \frac{dW_i(\tau)}{d\tau}\right\rangle$

$= 2\left\langle W_i(\tau), -\frac{dL}{dW_i}\right\rangle$

$\Rightarrow$ apply lemma

$\Rightarrow$ independent of i.