Clarke Differential

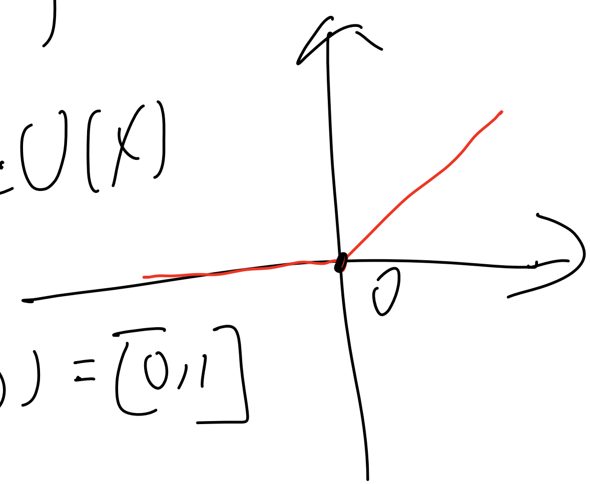
Subdifferential and Subgradient

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.

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

$g_t \in \partial_s f(x_t)$

$ReLU(x)$



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

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 & differentiable $\Rightarrow O\left(\frac{1}{T}\right)$

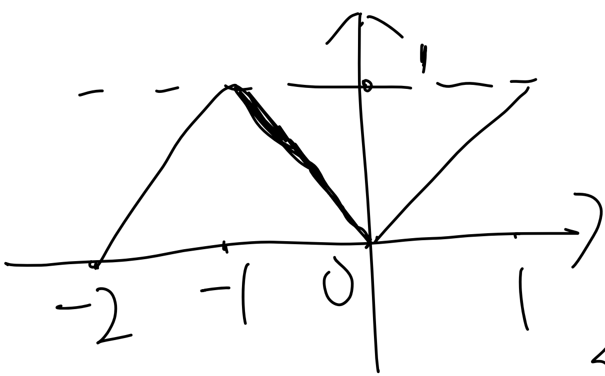
$\partial_s f(x) = \{\nabla f(x)\}$

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

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 opt 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 $x' = 1$

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

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

Clarke Differential

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

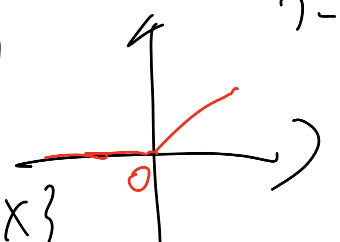
$$\partial f(x) \triangleq \text{conv}\left(\{s \in \mathbb{R}^d : \exists \{x_i\}_{i=1}^{\infty} \to x, \{\nabla f(x_i)\}_{i=1}^{\infty} \to s\}\right).$$

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

$\max\{0, x\}$



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

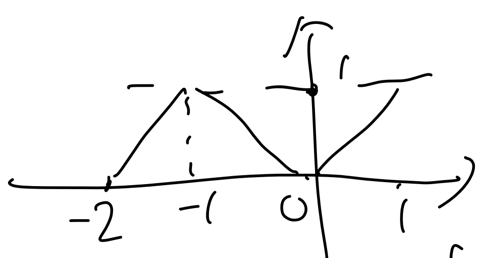
  $\nabla f(x_i) = 0$

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

  $\nabla f(x_i) = 1$

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

example:



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

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

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

$\Rightarrow g_t \in \partial f(x_t)$

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

When does Clarke differential exists

$\forall x'$
$|f(x') - f(x)|$
$\leq L \cdot \|x' - x\|$

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 $\rightarrow$ ∂f exists everywhere
- If f is convex $\Rightarrow$ $\partial f = \partial_s f$
- If f is differentiable $\Rightarrow$ $\partial f = \{\nabla f\}$

Positive Homogeneity

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

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

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

Positive Homogeneity

④ multi-layer ReLU

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

— for one-layer

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

— for all layers

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

$\Rightarrow$ degree $H+1$

Positive Homogeneity

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

$$\left\langle W_h, \frac{\partial f(x, W_1, \ldots, W_{H+1})}{\partial W_h} \right\rangle = f(x, W_1, \ldots, W_{H+1})$$

Pf: $A_h = \mathrm{diag}(\sigma'(W_h \sigma(\cdots \sigma(W_1 x) \cdots)))$

$\sigma' = 0$ or 1

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

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

$\Rightarrow$ 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 $\langle s, x \rangle = Lf(x)$.

$L=1$

W_M

Norm Preservation

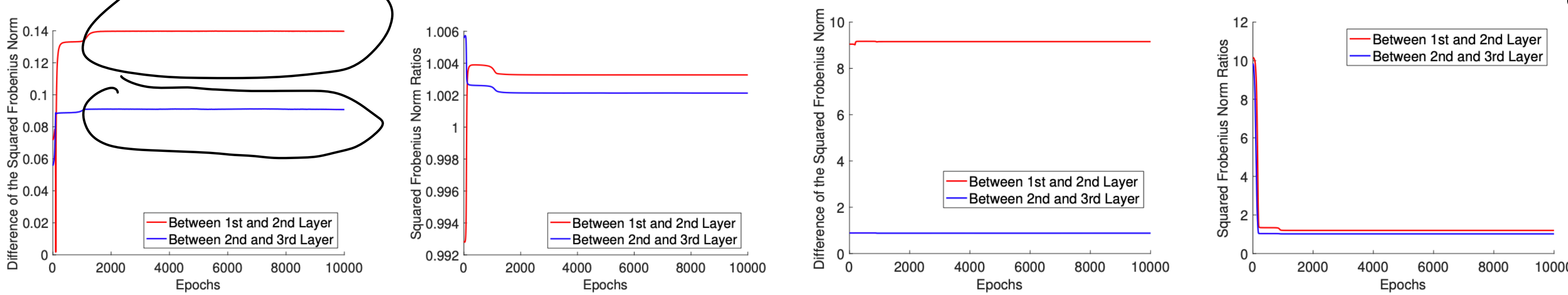
$f(x, W_1, W_2, W_3) = W_3\, \sigma(W_2\, \sigma(W_1 x))$

quadratic

$W_3 \times 10$

$W_2 / 10$

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



(a) Balanced initialization, squared norm differences.

(b) Balanced initialization, squared norm ratios.

(c) Unbalanced Initialization, squared norm differences.

(d) Unbalanced initialization, squared norm ratios.

$\ell(W) \leftarrow \|W_1\|_F^2 - \|W_2\|_F^2$

$\|W_2\|_F^2 - \|W_3\|_F^2$

suppose

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

if $W_1(0)$ & $W_2(0)$ small

Gradient flow and gradient inclusion

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

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

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

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 $(\cancel{i}\,h, \cancel{j}\,h') \in [H+1] \times [H+1]$, the gradient inclusion maintains: for all $t \geq 0$,

$$\frac{1}{2}\|W_h(t)\|_F^2 - \frac{1}{2}\|W_h(0)\|_F^2 = \frac{1}{2}\|W_{h'}(t)\|_F^2 - \frac{1}{2}\|W_{h'}(0)\|_F^2.$$

$$\Rightarrow \|W_h(t)\|_F^2 - \|W_{h'}(t)\|_F^2 = \|W_h(0)\|_F^2 - \|W_{h'}(0)\|_F^2$$

Pf: $\displaystyle\int \frac{d\|W_h(t)\|_F^2}{dt}\, dt$

Optimization Methods for Deep Learning

Gradient descent for non-convex optimization

Decsent Lemma: Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable, and $\|\nabla^2 f\|_2 \leq \beta$. Then setting the learning rate $\eta = 1/\beta$, and applying gradient descent, $x_{t+1} = x_t - \eta \nabla f(x_t)$, we have:

$$f(x_t) - f(x_{t+1}) \geq \frac{1}{2\beta} \|\nabla f(x_t)\|_2^2.$$

Pf: by Taylor Expansion & mean-value Theorem

$f(x+\Delta) = f(x) + \Delta^T \nabla f(x) + \frac{1}{2} \Delta^T \nabla^2 f(y) \Delta$ for some y

$\Delta^T \nabla^2 f(y) \Delta \leq \|\nabla^2 f(y)\|_2 \cdot \|\Delta\|_2^2 \leq \beta \|\Delta\|_2^2$

choose $\Delta = -\eta \nabla f(x_t)$

$f(x_{t+1}) \leq f(x_t) - \eta \|\nabla f(x_t)\|_2^2 + \frac{1}{2} \beta \eta^2 \|\nabla f(x_t)\|_2^2$

if $\eta = \frac{1}{\beta}$ $\quad = f(x_t) - \frac{1}{2\beta} \|\nabla f(x_t)\|_2^2$

Converging to stationary points

Theorem: In $T = O(\frac{\beta}{\epsilon^2})$ iterations, we have $\|\nabla f(x)\|_2 \leq \epsilon$.

Pf: $f(x_{t+1}) \leq f(x_t) - \frac{1/\beta}{2} \|\nabla f(x_t)\|_2^2$

sum over $t = 0, \ldots, T-1$

$$\sum_{t=1}^{T} f(x_t) \leq \sum_{t=0}^{T-1} f(x_t) - \frac{1/\beta}{2} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2$$

$$\Rightarrow \quad f(x_T) \leq f(x_0) - \frac{1/\beta}{2} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2$$

$$\Rightarrow \quad \underbrace{\sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2}_{} \leq \frac{2(f(x_0) - f(x_T))}{1/\beta}$$

$$T \min_{0 \leq t \leq T-1} \|\nabla f(x_t)\|_2^2 \leq \frac{2(f(x_0) - f(x_T))}{1/\beta}$$

$$\Rightarrow \quad \min_{0 \leq t \leq T-1} \|\nabla f(x_t)\|_2 \leq \sqrt{\frac{2(f(x_0) - f(x_T))}{1/\beta \cdot T}} \overset{\text{want}}{\leq} \varepsilon$$

$$\Rightarrow T = O\left(\frac{\beta}{\varepsilon^2}\right)$$