Clarke Differential

Clarke Differential

$$\boxed{\begin{aligned} x_{t+1} &\leftarrow x_t - \eta y_t \\ y_t &\in \partial f(x) \end{aligned}}$$

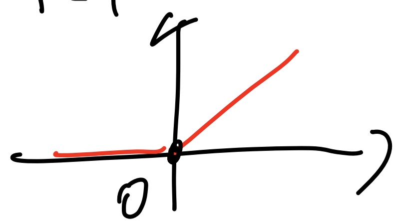
Definition: Given $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the Clark differential is defined as

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

The elements in the subdifferential set are subgradients.

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

example: ReLU



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

$\nabla f(x_i) = 0$

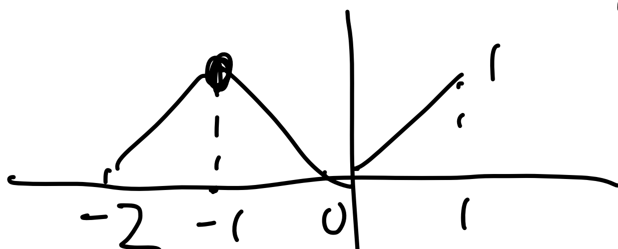
$\{x_i\}, 1, \frac{1}{2}, \dots \to 0$

$\nabla f(x_i) = 1$

$\Rightarrow \partial f(0) = \{\lambda_1 \cdot 0 + \lambda_2 \cdot 1,\ \lambda_1 + \lambda_2 = 1,\ \lambda_1, \lambda_2 \geq 0\}$

$= [0, 1]$

example

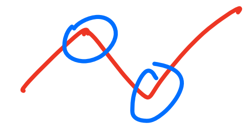


$\{x_i\}: -2, -1.5, \dots \to -1,\ \nabla f(x_i) = 1$

$\{x_i\}: 0, -0.5 \to -1,\ \nabla f(x_i) = -1$

$\Rightarrow \partial f(-1) = [-1, 1],$

When does Clarke differential exists



$|f(x)-f(x')| \le L\|x-x'\|, \quad x' \in \text{neighbor}(x)$

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.

(1) If f is locally Lip $\Rightarrow$ Clarke differential exists

$\Rightarrow$ NN with ReLU have clarke differential everywhere

- If f is convex $\Rightarrow \partial f = \partial_s f$
- If f is differentiable $\Rightarrow \partial f = \{\nabla f\}$

$\not\partial$ satisfies chain rule

Positive Homogeneity

motivate ReLU

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: $\prod_{j=1}^{d} x_j^{p_j}$, $\sum_{j=1}^{d} p_j = L$

$$\prod_{j=1}^{d} (\alpha x_j)^{p_j} = \alpha^{\sum p_j} \prod_{j=1}^{d} x_j^{p_j}$$

$$= \alpha^{L} \prod_{j=1}^{d} x_j^{p_j}$$

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

Positive Homogeneity

④ Multi-layer ReLU

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

for one-layer, degree-1

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

for all-layers

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

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

Positive Homogeneity

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

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

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

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

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

$$f(X, W_1, \ldots, W_{H+1}) = \underbrace{W_{H+1} A_H W_H \cdots A_1 W_1 X}_{\sigma(z) = \sigma'(z) \cdot z}$$

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

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

Positive Homogeneity and Clark Differential

⇒ clarke differential exist

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

Norm Preservation

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

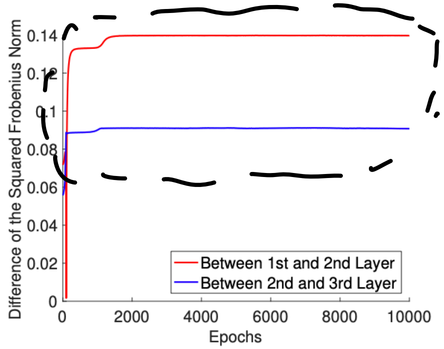
quadratic loss

$$\|W\|_F^2 = \sum_{i,j} W_{ij}^2$$

$W_3 \neq 0$

$W_2 \neq 0$

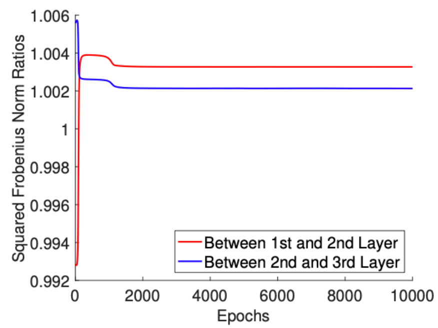
$$W_2(t+1) \leftarrow W_2(t) - \eta \frac{\partial}{\partial W_2}$$



(a) Balanced initialization, squared norm differences.

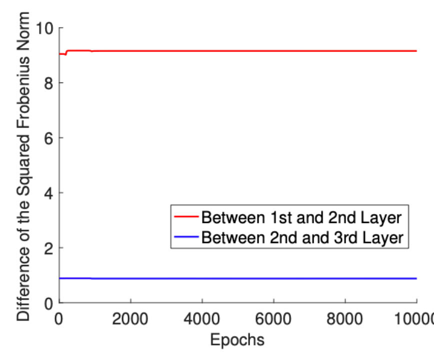
$\|W_1\|_F^2 - \|W_2\|_F^2$

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

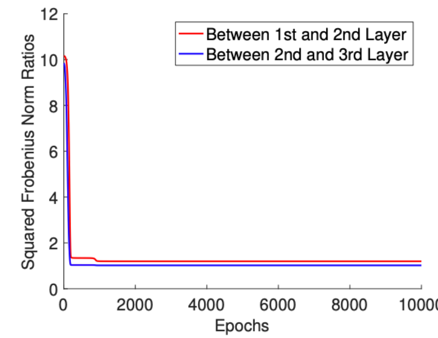


(b) Balanced initialization, squared norm ratios.

$$\frac{\|W_1\|_F^2}{\|W_2\|_F^2}$$



(c) Unbalanced Initialization, squared norm differences.



(d) Unbalanced initialization, squared norm ratios.

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

* algorithm regularization if $\|W_i(0)\|_F^2$ small for i $\Rightarrow$ $\|W_i(t)\|_F^2 \approx \|W_j(t)\|_F^2$

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 $t \geq 0$,

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

$$\Rightarrow \|W_i(t)\|_F^2 - \|W_j(t)\|_F^2 = \underbrace{\|W_i(0)\|_F^2 - \|W_j(0)\|_F^2}_{\text{small in the init}}$$

Pf: (1) $\frac{dW_i(t)}{dt}$ formula

(2) $\frac{1}{2}\|W_i(t)\|_F^2 - \frac{1}{2}\|W_i(0)\|_F^2 = \int_0^t \frac{d \frac{1}{2}\|W_i(\tau)\|_F^2}{d\tau} d\tau$

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 Thm

$$f(x+\delta) = f(x) + \delta^\top \nabla f(x) + \frac{1}{2} \delta^\top \nabla^2 f(y) \delta \quad \text{for some } y$$

(GD: $f(x+\delta) \approx f(x) + \delta^\top \nabla f(x) + \|\delta\|_2^2$

$\Rightarrow$ optimize $\delta \to -\nabla f(x)$)

$\delta^\top \nabla^2 f(y) \delta \leq \beta \cdot \|\delta\|_2^2$, choose $\delta = -\eta \nabla f(x)$

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

Converging to stationary points

approximate stationary point

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

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

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

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

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

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

$$\frac{\eta}{2}\cdot T \min_{0\le t\le T-1}\|\nabla f(x_t)\|_2^2 \le f(x_0) - \min_x f(x)$$

$$\Rightarrow \min_{0\le t\le T-1}\|\nabla f(x_t)\|_2 \le \sqrt{\frac{f(0) - \min_x f(x)}{\eta(T-1)}} = O(\epsilon)$$

Gradient Descent for Quadratic Functions

$x = 0$

Problem: $\min_x \frac{1}{2} x^\top A x$ with $A \in \mathbb{R}^{d \times d}$ being positive-definite.

Theorem: Let $\lambda_{\max}$ and $\lambda_{\min}$ be the largest and the smallest eigenvalues of A. If we set $\eta \leq \frac{1}{\lambda_{\max}}$, we have

$$\|x_t\|_2 \leq (1 - \eta\lambda_{\min})^t \|x_0\|_2$$

$$\|x_{t+1}\|_2 = \|x_t - \eta A x_t\|_2$$
$$= \|(I - \eta A) x_t\|_2$$
$$\leq \|I - \eta A\|_2 \|x_t\|_2$$
$$\leq (1 - \eta \lambda_{\min}) \|x_t\|_2$$
$$\leq (1 - \eta \lambda_{\min})^{t+1} \|x_0\|_2$$

To make $\|x_t\|_2 \leq \epsilon$

$\eta = \frac{1}{\lambda_{\max}}$

$\Rightarrow$ Need $\frac{\lambda_{\max}}{\lambda_{\min}} \log\left(\frac{1}{\epsilon}\right)$ iterations

$K = \frac{\lambda_{\max}}{\lambda_{\min}}$ condition number

$$A = \begin{pmatrix} 10 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

Momentum: Heavy-Ball Method (Polyak '64)

$V_0 = 0$

Problem: $\min_x f(x)$

Method: $v_{t+1} = -\nabla f(x_t) + \beta v_t$

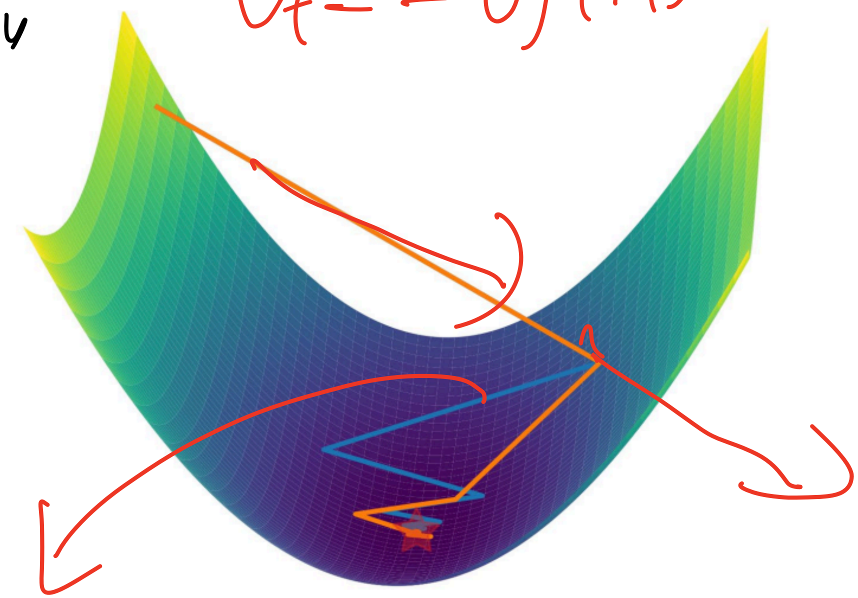
$x_{t+1} = x_t + \eta v_{t+1}$

for quadratic optimization

$O(\sqrt{\kappa} \log(\frac{1}{\epsilon}))$

v.s. $\kappa \log(\frac{1}{\epsilon})$

$v_t = -\nabla f(x_{t-1})$



$-\nabla f(x_t)$

Momentum: Nesterov Acceleration (Nesterov '89)

Problem: $\min_x f(x)$

Method: $v_{t+1} = -\nabla f(x_t + \beta v_t) + \beta v_t$

$x_{t+1} = x_t + \eta v_{t+1}$

$-\nabla f(x_t)$

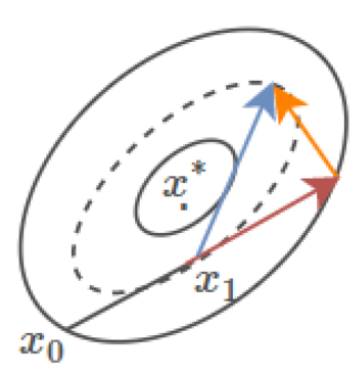
and

Theory

$\sqrt{\kappa} \log(\frac{1}{\varepsilon})$

for general strongly convex functions

Polyak's Momentum



Nesterov Momentum