CS 839: Theoretical Foundations of Deep Learning		Spring 2023
Lecture 9 Clarke Sub-differential and Positive Homogeneity		
Instructor: Yingyu Liang	*Date:*	*Scriber: Yang Guo*

1 Overview

In the previous lectures, we discussed how gradient descent (flow) algorithm induces an implicit bias, and converges to good optimum. However, our previous discussion requires the loss function to be differentiable, which is not true in some interesting cases such as ReLU networks, or the hinge loss. Thus, we will introduce Clarke sub-differential to handle differentiability. Also, we will introduce positive homogeneity which is an important property of typical ReLU networks.

2 Clarke Sub-differential

2.1 Sub-differential

Definition 1 (Sub-differential)**.** We define the sub-differential for function f as the following set:

$$\partial_s f(x) := \{v : \forall z, f(z) \geqslant f(x) + v^\top (z - x)\}$$

Remark 2. We note that:

1. If f is convex, $\partial_s f(x) \neq \varnothing$.
2. If f is convex, differentiable: $\partial_s f(x) = \{\nabla f(x)\}$.

2.2 Clarke Sub-differential

Definition 3 (Lipschitz Function)**.** A function $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz continuous if $\exists$ a constant L, s.t. $\forall x_1, x_2 \in \mathcal{X}$:

$$\|f(x_1) - f(x_2)\| \leqslant L\|x_1 - x_2\|$$

.

Definition 4 (Locally Lipschitz Function)**.** A function $f : \mathcal{X} \to \mathbb{R}$ is locally lipschitz if $\forall x \in \mathcal{X}$, $\exists$ a neighborhood $U(x)$ s.t. f is Lipschitz continuous on $U(x)$.

Definition 5 (Clarke Sub-differential)**.** We define the Clarke Sub-differential for a function f at $x \in \mathcal{X}$ as the following convex hull:

$$\partial f(x) := \text{conv}\{s : \exists x_i \to x, \nabla f(x_i) \to s\}$$

Remark 6. We note the following important properties:

1. If f is locally Lipschitz, $\forall x$, $\partial f(x)$ exists.
2. If f is convex, then $\forall x, \partial f(x) = \partial_s f(x)$.
3. If f is continuously differentiable at $x \iff \partial f(x) = \{\nabla f(x)\}$.

An example of locally lipschitz function is ReLU activation function. At the origin, $\lim_{x\to 0^+} \nabla f(x) = 1$, and $\lim_{x\to 0^-} \nabla f(x) = 0$. Thus, $\partial f(0) = \text{conv}\{0, 1\} = [0, 1]$

The importance of introducing Clarke sub-differential is to enable us performing chain rule for sub-gradient on non-smooth function.

2.3 Differential Inclusion Equation

Differential inclusion equation is an extension of ODE to function not differentiable.

Consider the example of gradient flow. For the differentiable loss function L, we can define the gradient flow as:

$$\dot{w}(t) = -\nabla L(w)$$

For L non differentiable, we obtain an corresponding differential inclusion equation:

$$\dot{w}(t) \in -\partial L(w)$$

Chain Rule Under some technical condition, like 0-minimal definability [2], we have chain rule: For a.e. $t \geqslant 0$, and $\forall v \in \partial L(w(t))$, we have

$$\frac{d}{dt} L(w(t)) = \langle v, \dot{w}(t) \rangle$$

Application of Chain Rule: Minimum Norm Path With the chain rule, we can have:

$$\dot{w}(t) = -\arg\min_v \{\|v\| : v \in \partial L(w(t))\}, \text{ for a.e. } t \geqslant 0$$

That is, the gradient flow is along the negative minimum norm sub-gradient. To see this, first note that $\partial L(w(t))$ is a compact set, so the minimum norm v exists; denote it as:

$$v^* := \arg\min_v \{\|v\| : v \in \partial L(w(t))\}.$$

Assume for contradiction that

$$\|\dot{w}(t)\| < \|v^*\|.$$

Apply the chain rule on $v = -\dot{w}(t)$, we have

$$\left|\frac{d}{dt} L(w(t))\right| = |\langle -\dot{w}(t), \dot{w}(t) \rangle| = \|\dot{w}(t)\|^2.$$

Apply the chain rule on $v = v^*$, we have

$$\left|\frac{d}{dt} L(w(t))\right| = |\langle v^*, \dot{w}(t) \rangle| \leqslant \|v^*\| \|\dot{w}(t)\| < \|\dot{w}(t)\|^2 = \left|\frac{d}{dt} L(w(t))\right|.$$

This is a contradiction. So we must have $\|\dot{w}(t)\| = \|v^*\|$.

Now, we will show that the minimum norm path will ensure non-increasing function value and convergence to local optimum:

$$\begin{aligned} L(w(t)) - L(w(0)) &= \int_0^t \frac{d}{ds} L(w(s)) ds \\ &= -\int_0^t \langle v, \dot{w}(s) \rangle ds \\ &\leqslant -\int_0^t \min\{\|v\|_2^2 : v \in \partial L(w(s))\} ds \\ &\leqslant -t \min_{s\in[0,t], v\in\partial L(w(s))} \|v\|^2 \end{aligned} \tag{1}$$

The first equality follows from the chain rule. The third line follows from the definition of $v, \dot{w}$. Re-arranging the last inequality gives:

$$\min_{s\in[0,t], v\in\partial L(w(s))} \|v\|^2 \leqslant \frac{L(w(0)) - L(w(t))}{t} \leqslant \frac{L(w(0))}{t}.$$

This implies for sufficiently large t, there exists some sub-differential with a small norm.

3 Positive Homogeneity

Definition 7 (Homogeneous)**.** A function g is L-homogeneous if

$$g(\alpha x) = \alpha^L g(x), \ \forall \alpha \geqslant 0$$

Note that this implies $g(0) = 0$ for homogeneous functions g.

Example 8. The ReLU activation function $\sigma(x) = \max(0, x)$ is 1-homogeneous, since $\sigma(\alpha x) = \max(0, \alpha x) = \alpha \max(0, x) = \alpha\sigma(x)$ for $\alpha \geqslant 0$.

Example 9. A monomial of degree L is L-homogeneous:

$$m(x) := \prod_i^d x_i^{P_i}$$

where $\sum_i P_i = L$, since $m(\alpha x) = \prod_i^d (\alpha x_i)^{P_i} == \alpha^{\sum_i P_i} \prod_i^d x_i^{P_i} = \alpha^L m(x)$ for $\alpha \geqslant 0$.

Example 10. A single layer of ReLU network is 1-homogeneous. (Viewing only one layer of parameter as function input). Recall the definition of L-layer ReLU network:

$$f(x; \theta) = f(x; W_1, \ldots, W_L) = W_L \sigma(W_{L-1} \sigma(\ldots \sigma(W_1 x)))$$

Since ReLU activation is 1-homogeneous,

$$\begin{aligned} f(x; W_1, \ldots, \alpha W_i, \ldots, W_L) &= W_L \sigma(W_{L-1} \sigma(\ldots \alpha W_i \sigma(\ldots \sigma(W_1 x)))) \\ &= \alpha W_L \sigma(W_{L-1} \sigma(\ldots W_i \sigma(\ldots \sigma(W_1 x)))) = \alpha f(x; \theta) \end{aligned} \tag{2}$$

Example 11. The entire L-layer ReLU network is L-homogeneous. (Viewing the parameters of all layers as the function input) Simply applying 1-homogeneous property to the parameter at each layer gives this property.

Clarke Sub-differential of L-layer ReLU Network

Definition 12 (Activation Matrix)**.** We define the activation matrix for i^{th} layer of the ReLU network as the following diagonal matrix:

$$A_i = \text{diag}(\sigma'(W_i\sigma(\dots))).$$

Note that $\sigma(x) = x\sigma'(x)$. Thus, we can rewrite the ReLU network as a matrix multiplication:

$$f(x;\theta) = W_L A_{L-1} \dots A_1 W_1 x.$$

And the gradient w.r.t the parameter of a single layer is given as:

$$\frac{d}{dW_i} f(x;\theta) = (W_L A_{L-1} \dots A_i)^\top (A_{i-1} W_{i-1} \dots A_1 W_1 x)^\top$$

And $\langle \frac{df(x;\theta)}{dw_i}, w_i \rangle = f(x;\theta)$. This equality follows from the cyclic property of trace.

Other Good Resources [3],[1] also provide very good reference.

References

[1] Quanquan Gu. Cs 269 foundations of deep learning.

[2] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

[3] Matus Telgarsky. Deep learning theory lecture notes.