CSE 543 / 599I - Deep Learning

Lecture 5: Clarke differential

April 12, 2022

Lecturer: Simon Du *Scribe: Prashant Rangarajan*

We continue with a discussion on optimization of neural networks. Often activation functions are not smooth (for example ReLU), in which case we can't use gradient descent directly. In this lecture we formalize the notion of the Clark Differential as a relaxation of the gradient.

1 Subdifferential and Subgradient

Definition 1.1. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the subdifferential set is defined as:

$$\partial_{\mathrm{s}} f(x) := \left\{ s \in \mathbb{R}^d : \forall x' \in \mathbb{R}^d, f(x') \geq f(x) + s^\top (x' - x) \right\}$$

The *subdifferential* is a set of tangents which lie below the predictor, each element of which is referred to as a *subgradient*.
For example, in the case of ReLU i.e. $f(x) = \max(0, x)$, we have $\partial_{\mathrm{s}} f(x) = [0, 1]$.

We can use these subgradients to perform *subgradient descent*:

$$\text{Problem: } \min_x f(x)$$
$$\text{Update Step: } x_{t+1} = x_t - \eta g_t \text{ where } g_t \in \partial_{\mathrm{s}} f(x_t)$$

Sometimes when the neural net is not convex the subgradient may not be well defined. Consider the function $f(x)$ depicted in figure 1. We try to obtain the set of subgradients at $x = -1$. Using the definition of subgradient, we can say that for all $s \in \mathbb{R}$,

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

Now, in the case when $x' = -2$, the above inequality reduces to $s \geq 1$, and for $x' = 1$, we get $s \leq 0$. Hence, there doesn't exist a subgradient at this point. So, to get some notion of this differential for such functions, we can relax the condition of requiring the inequality to hold for all x'. This gives rise to the *Clark differential.*

2 Clark Differential

Definition 2.1. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, for every x, the Clark differential $\partial f(x)$ is defined as:

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

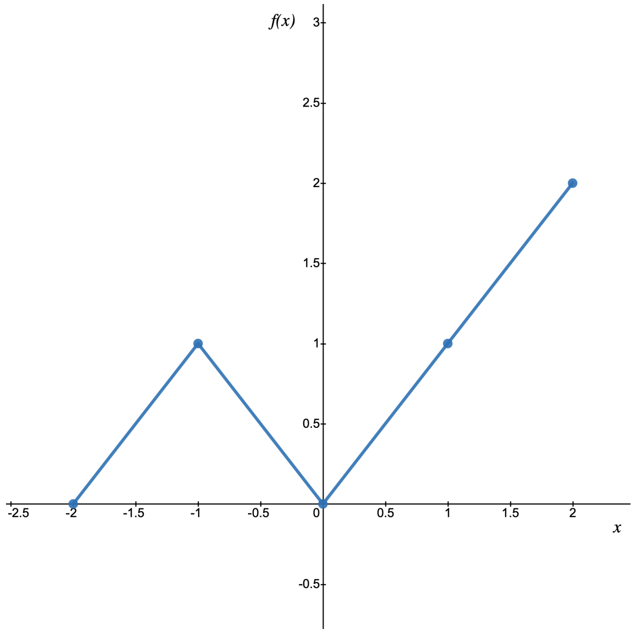

Figure 1: Non-convex function $f(x)$

Here conv(S) represents the convex hull of the set S:

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

For example, in the case of ReLU(x), at $x = 0$, we have a sequence $\{x_i\}_{i=1}^{\infty} = -1, \frac{-1}{2}, \ldots \to 0$ where $\nabla f(x_i) \to 0$, while there is another sequence $1, \frac{1}{2}, \ldots \to 0$ s.t. $\nabla f(x_i) \to 1$. Hence, $\partial f(x) = \text{conv}(\{0, 1\}) = [0, 1]$.
In the case of the function from figure 1, at $x = -1$, we have a sequence $\{x_i\}_{i=1}^{\infty} = -2, \frac{-3}{2}, \ldots \to -1$ where $\nabla f(x_i) \to -1$, while there is another sequence $0, \frac{-1}{2}, \ldots \to -1$ s.t. $\nabla f(x_i) \to -1$. Thus, $\partial f(-1) = [-1, 1]$.

Existence of Clark Differential To get sufficient conditions as to when the Clark Differential exists, we can use the following property.

Definition 2.2. A function $f : \mathbb{R}^d \to R$ is *locally Lipschitz* if $\forall x \in \mathbb{R}^d$, there exists a neighborhood S of x, such that f is Lipschitz in S. Hence, $\forall x' \in S$, $\exists L$ s.t. $|f(x) - f(x')| \leq L|x - x'|$.

Some key properties of the differential:

- If f is Locally Lipschitz, then ∂f exists everywhere.
- If f is convex, then $\partial f = \partial_{\text{s}} f$
- If f is differentiable, then $\partial f = \{\nabla f\}$

3 Positive Homogeneity

Positive Homogeneity is another useful tool that we can use in the analysis of gradients.

Definition 3.1. A function $g : \mathbb{R}^d \to \mathbb{R}$ is said to be positively homogeneous of degree L if $g(\alpha x) = \alpha^L g(x)$ for $\alpha \geq 0$.

Some examples:

- ReLU: $\sigma(\alpha x) = \alpha\sigma(x) \quad$ (1-homogeneous).
- Norm: $\|\alpha x\| = \alpha\|x\| \quad$ (1-homogeneous)
- Monomial of degree L i.e. of the form $\prod_{i=1}^d x_i^{p_i}$ where $\sum_{i=1}^d p_i = L$:

$$\prod_{i=1}^{d} (\alpha x_i)^{p_i} = \alpha^{\sum_i p_i} \prod_i x_i^{p_i} = \alpha^L \prod_i x_i^{p_i} \quad (L\text{-homogeneous})$$

- Consider a multi-layer ReLU network

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

where $x \in \mathbb{R}^d, W_1 \in \mathbb{R}^{m\times d}, W_2, \ldots W_H \in \mathbb{R}^{m\times m}, W_{H+1} \in \mathbb{R}^m$.

  - Layers of a ReLU network are 1-homogeneous in the parameters for that layer:

$$\begin{aligned}
&f(x; (W_1, \ldots, \alpha W_i, \ldots, W_{H+1})) \\
&= W_{H+1}\sigma(W_H\sigma(\ldots \alpha W_i\sigma(\ldots W_1 x \ldots) \ldots)) \\
&= \alpha W_{H+1}\sigma(W_H\sigma(\ldots W_i\sigma(\ldots W_1 x \ldots) \ldots)) \\
&= \alpha f(x; (W_1, \ldots, W_i, \ldots, W_{H+1})) = \alpha f(x; w)
\end{aligned}$$

  - The entire network is $(H + 1)$-homogeneous in the full set of parameters:

$$\begin{aligned}
&f(x; (\alpha W_1, \ldots, \alpha W_{H+1})) \\
&= \alpha W_{H+1}\sigma(\alpha W_H\sigma(\ldots\sigma(\alpha W_1 x) \ldots)) \\
&= \alpha^{H+1} W_{H+1}\sigma(W_H\sigma(\ldots\sigma(W_1 x) \ldots)) \\
&= \alpha^{H+1} f(x; w)
\end{aligned}$$

Special property of ReLU Networks For the ReLU network $f(x; w) = W_{H+1}\sigma_H(\cdots W_2\sigma_1(W_1 x))$, we get the following property $\forall h, 1 \leq h \leq H + 1$:

$$\left\langle W_h, \frac{\mathrm{d}}{\mathrm{d}W_h} f(x; w) \right\rangle = f(x; w)$$

So this inner product is independent of h.

Proof. Let A_h be a diagonal matrix with activations of the output after layer h on the diagonal:

$$A_h = \text{diag}\left(\sigma(W_h\sigma(\ldots\sigma(W_1x)\ldots))\right) \quad A_h \in \mathbb{R}^{m\times m}$$

We know that in the case of ReLU, $\sigma(z) = z\sigma'(z)$. Using this, we can rewrite the network as follows:

$$f(x;w) = W_{H+1}A_HW_HA_{H-1}\cdots A_1W_1x$$

The gradient of the network with respect to layer h is thus:

$$\frac{\text{d}}{\text{d}W_h}f(x;w) = (W_{H+1}A_H\cdots W_{h+1}A_h)^\top(A_{h-1}W_{h-1}\cdots W_1x)^\top$$

And so we can say that:

$$\begin{aligned}
&\left\langle W_h, \frac{\text{d}}{\text{d}W_h}f(x;w)\right\rangle \\
&\qquad = \left\langle W_h, (W_{H+1}A_H\cdots W_{h+1}A_h)^\top(A_{h-1}W_{h-1}\cdots W_1x)^\top\right\rangle \\
&\qquad = \text{tr}\left(W_h^\top(W_{H+1}A_H\cdots W_{h+1}A_h)^\top(A_{h-1}W_{h-1}\cdots W_1x)^\top\right) \\
&\qquad = \text{tr}\left((W_{H+1}A_H\cdots W_{h+1}A_h)^\top(W_hA_{h-1}W_{h-1}\cdots W_1x)^\top\right) \\
&\qquad = \text{tr}\left((W_hA_{h-1}W_{h-1}\cdots W_1x)^\top(W_{H+1}A_H\cdots W_{h+1}A_h)^\top\right) \\
&\qquad = \text{tr}\left(W_{H+1}A_H\cdots W_{h+1}A_hW_hA_{h-1}W_{h-1}\cdots W_1x\right) \\
&\qquad = f(x;w)
\end{aligned}$$

□

f is 1-homogeneous with respect to x (and L-homogeneous for w). We can also obtain a more general result in the case of an L-homogeneous function.

Lemma 3.2. *If $f:\mathbb{R}^d \to R$ is Locally Lipschitz and L-positively homogeneous, then for any $x \in \mathbb{R}^d$ and $s \in \partial f(x)$, we have $\langle s, x\rangle = Lf(x)$*

The proof is left as an exercise.

4 Norm Preservation

Positive homogeneity can be used to describe the dynamics of gradient descent. If predictions are positive homogeneous with respect to each layer, then gradient flow preserves norms of layers.

Gradient Inclusion Discrete-time dynamics can be complex. In order to simplify this, we can use continuous- time dynamics.
As $\eta \to 0$, we get a simpler version of the discrete dynamics known as *gradient flow* or an active gradient flow.

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

In the case of Clark differential, we use the notion of *Gradient inclusion*:

$$\frac{dx(t)}{dt} \in -\partial f(x(t))$$

Norm preservation by gradient inclusion In the idealized setting of the gradient flow, we can prove norm preservation.

Theorem 4.1. *(Simon S. Du, Hu, and Lee (2018))*
Suppose $\alpha > 0$, and $f(x;(W_1,\dots,\alpha W_i,\dots,W_{H+1})) = \alpha f(x;(W_1,\dots,W_i,\dots,W_{H+1}))$ i.e. the predictions are 1-positively homogeneous in each layer. Then, for $h, h' \in [H+1] \times [H+1]$, the gradient inclusion maintains for all $t \geq 0$:

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

Proof. Assume that the loss function used to update the weights is of the form $\ell(f(x;(W_1 \dots W_{H+1})), y)$. Then, we can say using gradient inclusion that:

$$\frac{dW_h(t)}{dt} = -\nabla \ell(W_h(t)) = -\ell'(f(x;w),y) \cdot \frac{\partial f}{\partial W_h}$$

using Chain Rule. Thus, we can say that:

$$\begin{aligned}
\frac{1}{2}\|W_h(t)\|^2 - \frac{1}{2}\|W_h(0)\|^2 &= \int_0^t \frac{d}{d\tau}\|W_h(\tau)\|^2 d\tau \\
&= \int_0^t \left\langle W_h, \frac{dW_h(\tau)}{d\tau} \right\rangle d\tau \quad \text{(Chain rule)} \\
&= \int_0^t \left\langle W_h, -\ell'(f(x;w),y) \cdot \frac{\partial f}{\partial W_h} \right\rangle d\tau \\
&= \int_0^t -\ell'(f(x;w),y) \left\langle W_h, \frac{\partial f}{\partial W_h} \right\rangle d\tau \\
&= \int_0^t -\ell'(f(x;w),y) f(x;w) d\tau
\end{aligned}$$

where we used the Lemma 3.2 in the last step. Hence, the RHS independent of h. Thus, $\frac{1}{2}\|W_h(t)\|^2 - \frac{1}{2}\|W_h(0)\|^2$ is a constant for all $h \in [H+1]$, and so the result holds. □