14. ReLU net optimization & Implicit bias - 1

Last class

- Analyzed the behavior of neural nets with smooth activations
  - Considered a linearized version $u(t)$, trained with the gradient flow
  - Assumed that initial parameters have been scaled: $f \mapsto \alpha f$

  - Risk convergence: $\hat{R}(\alpha \cdot f(w(t))) \leq \hat{R}_0 \cdot \exp(-c\alpha^2 t)$
  - Parameter convergence: $\|w(t) - w(0)\| \leq c\sqrt{\hat{R}_0}/\alpha$

This week

- Optimization properties of neural networks with ReLU activations
  - Basic tool:
    - Clarke subdifferential (recap)
    - Locally Lipschitz
    - Positive homogeneity
  - Key results:
    - Norm preservation & automatic balancing
    - Risk convergence
    - Margin maximization

Tools

Subdifferential set

- **Problem.** ReLU is non-differentiable
  - Less problem with GD
    - Differentiable almost everywhere
  - More problem with GF

Definition (subdifferential set).

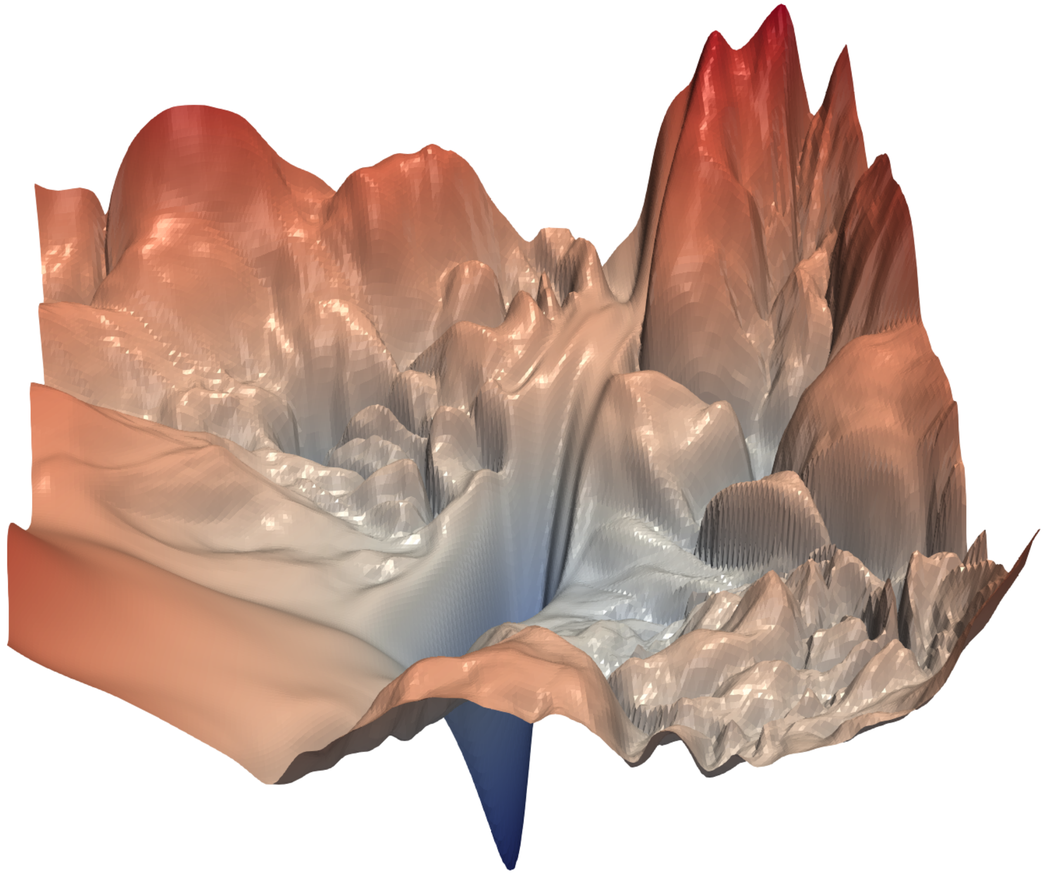
The subdifferential set ∂_s is the set of tangents which lie below the function

$$\partial_s \hat{R}(w) := \left\{ s \in \mathbb{R}^p \;:\; \hat{R}(w') \geq \hat{R}(w) + s^\top (w' - w), \quad \forall w' \right\}$$

- The elements are called “subgradients”
- If $\hat{R}$ is convex, non-empty everywhere
- If $\hat{R}$ is convex and $\nabla \hat{R}$ exists at w, then we have $\partial_s \hat{R}(w) = \{ \nabla \hat{R}(w) \}$

Clarke differential

- **Problem.** If $\hat{R}$ is not convex, then $\partial_s \hat{R}(w)$ can be empty
  - Usually the case with neural nets



Clarke differential

Definition (Clarke differential).

The Clarke differential is a set given as

$$\partial \hat{R}(w) := \operatorname{conv}\left(\left\{ s \in \mathbb{R}^p : \exists w_i \to w, \nabla \hat{R}(w_i) \to s \right\} \right)$$

- Limits of convergent sequences — directional derivatives
- Defined even for nonconvex functions

- Can be used to make our gradient flow well-defined.

- **Question.** Can we guarantee that Clarke differential always exist?

Locally Lipschitz

- To do this, we need a relaxed version of the Lipschitz property
    - From global to local

Definition (Locally Lipschitz).

A function $f(\cdot)$ is locally Lipschitz whenever for every point x, there exists a neighborhood $S \supset \{x\}$ such that $f(\cdot)$ is Lipschitz when restricted to S

- If $\hat{R}$ is locally Lipschitz, then $\partial\hat{R}$ exists everywhere

- **Exercise.** Can you think of any function that is:
    - Not Lipschitz, but locally Lipschitz?
    - Continuous but not locally Lipschitz?

Gradient flow, relaxed

- Now, consider a new gradient flow equation, where $=$ is replaced with $\in$

$$\dot{w}(t) \in -\partial \hat{R}(w(t)), \quad \text{for a.e. } t \geq 0$$

- With some additional assumptions, we have useful properties:
  - **Chain rule.** For a.e. $t \geq 0$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{R}(w(t)) = -\langle v, \dot{w}(t)\rangle, \qquad \forall v \in \partial \hat{R}(w(t))$$

  - **Minimum norm path.** For a.e. $t \geq 0$, we can let

$$\dot{w}(t) = -\operatorname{argmin}\{\|v\| \mid v \in \partial \hat{R}(w(t))\}$$

Stationarity

- From the minimum norm path property, we have

$$\begin{aligned}\hat{R}(w(t)) - \hat{R}(w(0)) &= \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\hat{R}(w(s))\,\mathrm{d}s \\ &= -\int_0^t \min\left\{ \|v\|^2 \;\middle|\; v \in \partial\hat{R}(w(s)) \right\} \mathrm{d}s\end{aligned}$$

- This leads to a nice stationarity property

$$\min_{s\in[0,t]} \min_{v\in\partial\hat{R}(w(s))} \|v\|^2 \leq \frac{\hat{R}(w(0)) - \hat{R}(w(t))}{t}$$

Positive homogeneity

- For ReLU, there is a useful property

Definition (Positive homogeneity).

A function $g(\cdot)$ is positive homogeneous with degree L, whenever

$$g(\alpha \cdot x) = \alpha^L \cdot g(x), \qquad \forall \alpha \geq 0$$

- For ReLU, holds with $L = 1$
- For degree-k monomials, holds with $L = k$

- Quite useful for neural nets

Norm preservation

Clarke differential of ReLU nets

- Now, think about the Clarke differential of ReLU nets

- Consider a ReLU net:

$$x \mapsto W_L \sigma\big(W_{L-1}\sigma\big(\cdots\sigma\big(W_1 x\big)\big)\cdots\big)$$

  - As a function of x, this is 1-homogeneous and piecewise affine
  - As a function of $w = (W_1, \ldots, W_L)$, this is L-homogeneous and piecewise polynomial

  - To compute the Clarke differential on a piece, we can:
    - Compute the gradient for a piece and all adjacent pieces
    - Take a convex hull, on the boundaries

Activation matrix

- Consider an activation matrix, which indicates where a ReLU neuron is activated or not

- Let A_i be a diagonal matrix, with layer i activations on the diagonal:

$$A_i = \text{diag}\left(\sigma'(W_i\sigma(W_{i-1}\cdots(W_1x))\cdots)\right)$$

  - Diagonal elements are either 1 or 0
    - 1 denotes “active”
    - 0 denotes “zero-ed out”
  - Dependent on x

- Then, we have

$$W_i\sigma(\cdots\sigma(W_1x)\cdots) = W_iA_{i-1}W_{i-1}\cdots A_1W_1x$$

  - Why?

Activation matrix

- The activation matrix is quite useful — allows us to drop σ

- Function output:

$$f(x; w) = W_L A_{L-1} W_{L-1} A_{L-2} \cdots A_1 W_1 x$$

- Gradient w.r.t. layer i

$$\frac{\mathrm{d}}{\mathrm{d}W_i} f(x; w) = (W_L A_{L-1} \cdots W_{i+1} A_i)^\top (A_{i-1} W_{i-1} \cdots W_1 x)^\top$$

Weight-gradient preservation

- We have a weight-gradient preservation property:

$$\begin{aligned}
\left\langle W_i, \frac{\mathrm{d}}{\mathrm{d}W_i} f(x; w) \right\rangle &= \left\langle W_i, (W_L A_{L-1} \cdots W_{i+1} A_i)^\top (A_{i-1} W_{i-1} \cdots W_1 x)^\top \right\rangle \\
&= \mathrm{tr}\left(W_i^\top (W_L A_{L-1} \cdots W_{i+1} A_i)^\top (A_{i-1} W_{i-1} \cdots W_1 x)^\top \right) \\
&= \mathrm{tr}\left((W_L A_{L-1} \cdots W_{i+1} A_i)^\top (W_i A_{i-1} W_{i-1} \cdots W_1 x)^\top \right) \\
&= \mathrm{tr}\left((W_i A_{i-1} W_{i-1} \cdots W_1 x)^\top (W_L A_{L-1} \cdots W_{i+1} A_i)^\top \right) \\
&= \mathrm{tr}\left(W_L A_{L-1} \cdots W_{i+1} A_i W_i A_{i-1} W_{i-1} \cdots W_1 x \right) \\
&= f(x; w)
\end{aligned}$$

- That is, the Frobenius product of weights and gradients are same over all layers

Weight-gradient preservation

- More generally, we have the following result

Lemma 9.2.

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz and L-positively homogeneous.

Then, for any $w \in \mathbb{R}^d$ and $s \in \partial f(w)$, we have

$$\langle s, w \rangle = L \cdot f(w)$$

- Select w differently — W_1, W_2, ...
    - The right-hand side remains the same

Proof sketch

For any $w \in \mathbb{R}^d$ and $s \in \partial f(w)$, we have

$$\langle s, w \rangle = L \cdot f(w)$$

- Suppose that $w \neq 0$
  - Otherwise, holds trivially

- Now, we have two cases:
  - Case 1. $f(\cdot)$ is differentiable at w
  - Case 2. $f(\cdot)$ is non-differentiable at w

Proof sketch

For any $w \in \mathbb{R}^d$ and $s \in \partial f(w)$, we have

$$\langle s, w \rangle = L \cdot f(w)$$

- **Case 1.** $f(\,\cdot\,)$ is differentiable at w
- We have:

$$\lim_{\delta \to 0^+} \frac{f(w + \delta w) - f(w) - \langle \nabla f(w), \delta w \rangle}{\delta \|w\|} = 0$$

- Using the homogeneity, we have

$$-\langle \nabla f(w), w \rangle + \lim_{\delta \to 0^+} \frac{((1+\delta)^L - 1) f(w)}{\delta} = 0$$

- Rearranging and evaluating the limit, we get

$$\langle s, w \rangle = L \cdot f(w)$$

Proof sketch

For any $w \in \mathbb{R}^d$ and $s \in \partial f(w)$, we have

$$\langle s, w \rangle = L \cdot f(w)$$

- **Case 2.** $f(\cdot)$ is non-differentiable
- Use the definition of Clarke differential!
  - If s is a limiting point (i.e., not generated by convex combination), then take a sequence (w_i) such that

$$\lim_{i \to \infty} w_i = w, \qquad \lim_{i \to \infty} \nabla \hat{R}(w_i) \to s$$

  - Then, we have

$$\langle w, s \rangle = \lim_{i \to \infty} \langle w_i, \nabla f(w_i) \rangle = \lim_{i \to \infty} L \cdot f(w_i) = L \cdot f(w)$$

  - If $s = \sum \alpha_i s_i$ is generated by a convex combination, then take the limit separately.

Remark

- Roughly put, the lemma says that:

  “⟨weight, gradient⟩ is constant over the layers”

  - Large weight = Small gradients
  - Small weight = Large gradients

- This leads to certain balancing between layers!

Automatic balancing

Automatic balancing

- Concretely, we have the following lemma

Lemma 9.3. (Du et al., 2018).

Suppose that predictions are 1-positive homogeneous for each layer.

Then, for every pair of layers, the gradient flow satisfies

$$\|W_i(t)\|^2 - \|W_i(0)\|^2 = \|W_j(t)\|^2 - \|W_j(0)\|^2$$

- This is our first example of implicit regularization / bias
  - GD has imposed some conditions on the solutions we can search for
  - Thus, we are not utilizing the full parameter space

$$\min_{w \in \mathscr{W}} \hat{R}(w) \longrightarrow \min_{w \in \mathscr{W}_{\text{gd}} \subseteq \mathscr{W}} \hat{R}(w)$$

Proof sketch

- For simplicity, assume that the risk is of form:

$$\hat{R}(w) = \mathbb{E}_k \ell(y_k \cdot f(x_k; w))$$

  - Define the shorthand notation $\ell'_k(s) := y_k \cdot \ell'(y_k \cdot f(x_k; w(s)))$

- Then, we have:

$$\begin{aligned}
\|W_i(t)\|^2 - \|W_i(0)\|^2 = \int_0^t \frac{\mathrm{d}}{\mathrm{d}s} \|W_i(s)\|^2 \,\mathrm{d}s &= 2 \cdot \int_0^t \langle W_i(s), \dot{W}_i(s) \rangle \,\mathrm{d}s \\
&= 2 \cdot \int_0^t \left\langle W_i(s), -\mathbb{E}_k \ell'_k(s) \frac{\mathrm{d}f(x_k; w)}{\mathrm{d}W_i(s)} \right\rangle \mathrm{d}s \\
&= 2 \cdot \int_0^t \mathbb{E}_k \ell'_k(s) \left\langle W_i(s), -\frac{\mathrm{d}f(x_k; w)}{\mathrm{d}W_i(s)} \right\rangle \mathrm{d}s \\
&= 2 \cdot \int_0^t \mathbb{E}_k \ell'_k(s) f(x_k; w) \,\mathrm{d}s \qquad \text{<- independent of i}
\end{aligned}$$

Next up

- Risk convergence