15. ReLU net optimization & Implicit bias - 2

Settings

- For a warm-up, we now consider a two-layer ReLU net + logistic regression
  - Assume that only the first layer is trainable

$$f(x; w) = \frac{1}{\sqrt{m}} \sum_j a_j \sigma(\langle x, w_j \rangle), \qquad a_j \in \{\pm 1\}, \quad \|x\| \le 1$$

  - Let $W_s \in \mathbb{R}^{m \times d}$ denote the parameters at time s

  - Then, we have:

$$\nabla f(x; W) = \frac{1}{\sqrt{m}} \cdot \begin{bmatrix} a_1 \sigma'(w_1^\top x) x \\ \cdots \\ a_m \sigma'(w_m^\top x) x \end{bmatrix}$$

$$\left\| \nabla f(x; W) \right\|_F^2 = \frac{1}{m} \sum_{j=1}^m \left\| \sigma'(w_j^\top x) x \right\|_2^2 \le \frac{1}{m} \sum_{j=1}^m \|x\|_2^2 \le 1$$

Settings

- We perform the **gradient descent** with the logistic loss, i.e.,

$$\ell(z) = \log(1 + \exp(-z))$$

$$\ell'(z) = \frac{-\exp(-z)}{1 + \exp(-z)} \qquad \in (-1,0)$$

$$\hat{R}(W) := \frac{1}{n} \sum_{k=1}^{n} \ell\left(y_k \cdot f(x_k; W)\right)$$

- Here is a useful fact:

$$|\ell'(z)| = -\ell'(z) \leq \ell(z)$$

Settings

- From the useful fact, we have:

$$\nabla \hat{R}(W) = \frac{1}{n} \sum_{k=1}^{n} \ell'\left(y_k \cdot f(x_k; W)\right) \cdot y_k \cdot \nabla f(x_k; W)$$

$$\begin{aligned} \left\| \nabla \hat{R}(W) \right\|_F &\leq \frac{1}{n} \sum_{k=1}^{n} \left| \ell'\left(y_k \cdot f(x_k; W)\right) \right| \cdot \left\| y_k \cdot \nabla f(x_k; W) \right\|_F \\ &\leq \frac{1}{n} \sum_{k=1}^{n} \left| \ell'\left(y_k \cdot f(x_k; W)\right) \right| \\ &\leq \min\{1, \hat{R}(W)\} \end{aligned}$$

Risk convergence

- Using these, we can prove the following lemma

Lemma 9.4. (Ji & Telgarsky, 2019)

Let $\eta \leq 1$. Then, for any Z, we have

$$\sum_{i<t} \frac{1}{t} \hat{R}(W_i) - \sum_{i<t} \frac{2}{t} \hat{R}^{(i)}(Z) \leq \frac{\|W_0 - Z\|_F^2 - \|W_t - Z\|_F^2}{\eta \cdot t}$$

where

$$\hat{R}^{(i)}(Z) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_k \cdot \langle Z, \nabla f(x_k; W_i) \rangle)$$

- **Note.** We have $\hat{R}^{(i)}(W_i) = \hat{R}(W_i)$
  - We have $\hat{R}^{(i)}(Z) \approx \hat{R}(Z)$, if W_i and Z have similar activations
    (thus, for convergence, we need to prove that similar activation happens often)

Proof sketch

$$\sum_{i<t} \frac{1}{t} \hat{R}(W_i) - \sum_{i<t} \frac{2}{t} \hat{R}^{(i)}(Z) \leq \frac{\|W_0 - Z\|_F^2 - \|W_t - Z\|_F^2}{\eta \cdot t}$$

- Begin from the usual decomposition

$$\|W_{i+1} - Z\|_F^2 = \|W_i - Z\|_F^2 - 2\eta \langle \nabla \hat{R}(W_i), W_i - Z \rangle + \eta^2 \|\nabla \hat{R}(W_i)\|_F^2$$

- **Third term.** Proceed as

$$\|\nabla \hat{R}(W_i)\|_F^2 \leq \left(\min\{1, \hat{R}(W_i)\} \right)^2 \leq \hat{R}(W_i)$$

Proof sketch

$$\sum_{i<t} \frac{1}{t} \hat{R}(W_i) - \sum_{i<t} \frac{2}{t} \hat{R}^{(i)}(Z) \leq \frac{\|W_0 - Z\|_F^2 - \|W_t - Z\|_F^2}{\eta \cdot t}$$

$$\|W_{i+1} - Z\|_F^2 \leq \|W_i - Z\|_F^2 - 2\eta \langle \nabla \hat{R}(W_i), W_i - Z \rangle + \eta^2 \hat{R}(W_i)$$

- **Second term.** Proceed as

$$\begin{aligned}
\langle \nabla \hat{R}(W_i), Z - W_i \rangle &= \frac{1}{n} \sum_{k=1}^{n} y_k \cdot \ell'(y_k \cdot f(x_k; W_i)) \cdot \langle \nabla f(x_k; W_i), Z - W_i \rangle \\
&= \frac{1}{n} \sum_{k=1}^{n} \ell'(y_k \cdot f(x_k; W_i)) \cdot \Big(\langle y_k \cdot \nabla f(x_k; W_i), Z \rangle - y_k \cdot f(x_k; W_i) \Big) \\
&\leq \frac{1}{n} \sum_{k=1}^{n} \ell\Big(\langle y_k \cdot \nabla f(x_k; W_i), Z \rangle \Big) - \ell\Big(y_k \cdot f(x_k; W_i) \Big) \\
&\leq \hat{R}^{(i)}(Z) - \hat{R}(W_i)
\end{aligned}$$

Proof sketch

$$\sum_{i<t} \frac{1}{t} \hat{R}(W_i) - \sum_{i<t} \frac{2}{t} \hat{R}^{(i)}(Z) \leq \frac{\|W_0 - Z\|_F^2 - \|W_t - Z\|_F^2}{\eta \cdot t}$$

$$\|W_{i+1} - Z\|_F^2 \leq \|W_i - Z\|_F^2 + 2\eta \hat{R}^{(i)}(Z) - 2\eta \hat{R}(W_i) + \eta^2 \hat{R}(W_i)$$

- Reordering, we get

$$(2 - \eta)\hat{R}(W_i) - 2\hat{R}^{(i)}(Z) \leq \frac{\|W_i - Z\|_F^2 - \|W_{i+1} - Z\|_F^2}{\eta}$$

- Use the fact that $1 \leq 2 - \eta$, and telescope to get the claim

Margin maximization: a primer

Setting

- Consider a slightly different setup: Binary classification with separable data

- **Classifier.** The function $f(x; w)$ is locally Lipschitz and L-homogeneous
  - e.g., linear model $x \mapsto \langle w, x \rangle$ (followed by sgn($\cdot$) for classification)

- **Loss.** We use the exponential loss

$$\ell(y \cdot f(x; w)) = \exp(-y \cdot f(x; w))$$

  - Margin mapping.

$$m_i(w) := y_i \cdot f(x; w)$$

  - (Unnormalized) risk.

$$\mathscr{L}(w) := \sum_i \ell(m_i(w))$$

Setting

- We assume separable data, i.e., one can perfectly classify the training data

Definition (linearly separable).

A dataset is linearly separable if there exists $w \in \mathbb{R}^d$ such that

$$\min_{i \in [n]} y_i \cdot \langle w, x_i \rangle > 0$$

Definition ($\vec{m}$-separable).

A dataset is $\vec{m}$-separable if there exists $w \in \mathbb{R}^d$ such that

$$\min_{i \in [n]} m_i(w) > 0$$

- **Question.** For separable data, is there any way to achieve zero training risk?

Infimum not attained

- The answer is no

Proposition 10.1.

Suppose that $f(x; w)$ is L-homogeneous in w, and there exists $\hat{w}$ such that $\mathscr{L}(\hat{w}) < \ell(0)$.

Then, we have $\inf_w \mathscr{L}(w) = 0$ and the infimum is not attained.

- **Proof idea.** One can always do better by increasing the magnitude.

Proof sketch

- First, we have:

$$\ell(0) > \mathscr{L}(\hat{w}) = \sum_{i=1}^{n} \ell(m_i(\hat{w})) > \max_{i \in [n]} \ell(m_i(\hat{w}))$$

  - Applying ℓ^{-1} on both sides, we get $0 < \min_{i \in [n]} m_i(\hat{w})$

- Then, we proceed as

$$\begin{aligned} 0 \leq \inf_{w} \mathscr{L}(\hat{w}) &\leq \limsup_{c \to \infty} \mathscr{L}(c\hat{w}) \\ &= \sum_{i=1}^{n} \limsup_{c \to \infty} \ell(m_i(c\hat{w})) \\ &= \sum_{i=1}^{n} \limsup_{c \to \infty} \ell(c^L m_i(\hat{w})) = 0 \end{aligned}$$

Margin maximization

- Now, we'll focus on showing the following claim:

GD finds (or is implicitly biased toward) the max-margin predictor

- **(Linear) Max-margin predictor.**

$$\bar{u} = \operatorname{argmax}_{\|w\|=1} \min_{i \in [n]} y_i \cdot \langle w, x_i \rangle$$

  - In other words, we do not need a hard-margin SVM

- **Question.** Why do we need a norm-1 constraint?
- **Answer.** To provide a fair comparison:

$$m_i(w) = \|w\|^L \cdot m_i\left(\frac{w}{\|w\|}\right)$$

Margins

Definition (Margins).

The margin, maximum margin, and the smooth margin is defined as:

$$\gamma(w) = \min_{i \in [n]} m_i\left(\frac{w}{\|w\|}\right) = \frac{1}{\|w\|^L} \min_{i \in [n]} m_i(w)$$

$$\bar{\gamma} := \max_{\|w\|=1} \gamma(w), \qquad \tilde{\gamma}(w) := \frac{\ell^{-1}(\mathscr{L}(w))}{\|w\|^L}$$

- **Remark.** The smooth margin gives us a nice sandwich bound for margin.

$$\ell^{-1}(\mathscr{L}(w)) \leq \ell^{-1}(\max_i \ell(m_i(w))) = \min_{i \in [n]} m_i(w)$$

$$\min_{i \in [n]} m_i(w) \leq \ell^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \ell(m_i(w))\right) = \ell^{-1}(\mathscr{L}(w)) + \log(n)$$

Properties

- Here are some basic properties of margins

Proposition 10.2.

Suppose that the data is $\overrightarrow{m}$-separable. Then,

- $\bar{\gamma}$ is well-defined (i.e., maximum attained)
- For any $w \neq 0$, we have

$$\lim_{c \to \infty} \tilde{\gamma}(cw) = \gamma(w)$$

  - In particular, for $\hat{w}$ that achieves $\bar{\gamma}$, we have $\lim_{c \to \infty} \tilde{\gamma}(c\hat{w}) = \bar{\gamma}$

- **Proof idea.** Invoke continuity and the definition of margins

Gradient flow

Lemma 10.1.

Consider a linearly separable data, and a linear classifier with $w(0) = \mathbf{0}$.

Also, suppose that $\max_{i \in [n]} \|x_i\| \leq 1$.

Then, we have

$$\mathscr{L}(w(t)) \leq \frac{1 + \log(2nt\bar{\gamma}^2)}{2t\bar{\gamma}^2}$$

$$\|w(t)\| \geq \log(2tn\bar{\gamma}^2) - \log(1 + \log(2tn\bar{\gamma}^2))$$

- **Remark.**
  - Risk decays as $\log t/t$
  - Parameter grows as $\log t$
  - No discussions on implicit bias yet ;(

Proof sketch

- First, observe that $\mathscr{L}$ is convex, as it is a summation of n convex functions

- Also, recall that our Theorem 7.4., which states that for convex risk we have: For any $z \in \mathbb{R}^d$

$$\mathscr{L}(w(t)) \leq \mathscr{L}(z) + \frac{1}{2t}\left(\|w(0) - z\|^2 - \|w(t) - z\|^2\right)$$

- **Claim 1.** Plug in $z = \log(c) \cdot \bar{u}/\bar{\gamma}$, to get:

$$\mathscr{L}(w(t)) \leq \mathscr{L}(z) + \frac{\|z\|^2}{2t} = \sum_{i=1}^{n} \ell\left(\log c \cdot \frac{m_i(\bar{u})}{\bar{\gamma}}\right) + \frac{\log^2(c)}{2t\bar{\gamma}^2}$$

$$\leq \sum_{i=1}^{n} \exp\left(-\log c\right) + \frac{\log^2(c)}{2t\bar{\gamma}^2} = \frac{n}{c} + \frac{\log^2(c)}{2t\bar{\gamma}^2}$$

  - Plug in $c = 2tn\bar{\gamma}^2$ to get the first bound.

Proof sketch

- **Claim 2.**
  - First, note that as $\max_{i \in [n]} \|x_i\| \leq 1$, we have

$$\|w(t)\| \geq y_i \langle w(t), x_i \rangle = m_i(w(t)), \qquad \forall i \in [n]$$

  - Thus, we have

$$\ell(\|w_t\|) \leq \ell\left(\max_{i \in [n]} m_i(w(t)\right) = \min_{i \in [n]} \ell\left(m_i(w(t)\right) \leq \frac{1}{n} \cdot \mathscr{L}(w(t))$$

$$\leq \frac{1 + \log(2nt\bar{\gamma}^2)}{2nt\bar{\gamma}^2}$$

  - Applying ℓ^{-1} on both sides, we get what we want.

Margin maximization

Want to show

- Now, we want to prove, either
  - **Parameter convergence.**

$$\lim_{t\to\infty} w_t \to \text{max margin solution}$$

  - **Margin convergence.**

$$\lim_{t\to\infty} \min_{i\in[n]} m_i(w_t) = \text{max margin}$$

- **Problem.** We know that
  - w_t diverges to infinity
  - $\min_{i\in[n]} m_i(w_t)$ diverges to infinity

Want to show

- **Idea.** Consider the normalized quantities
  - Normalized weight

$$w_t / \|w_t\|$$

  - Normalized max margin

$$\bar{u} = \arg \max_{\|w\|=1} \min_{i \in [N]} y_i \cdot f(x_i; w)$$

- Then, we can show either
  - $\lim_{t \to \infty} \frac{w_t}{\|w_t\|} = \bar{u}$
  - $\lim_{w \to \infty} \min_{i \in [n]} m_i \left(\frac{w_t}{\|w_t\|} \right) = \min_{i \in [n]} m_i(\bar{u})$ <— We'll show this!

Convergence of normalized margins

- Recall the definition of margins:
  - (Normalized) Margin

$$\gamma(w) = \min_{i \in [n]} m_i \left(\frac{w}{\|w\|} \right)$$

  - Normalized max margin

$$\bar{\gamma} := \max_{\|w\|=1} \gamma(w)$$

- We want to show that

$$\lim_{t \to \infty} \gamma(w_t) = \bar{\gamma}$$

- **Problem.** Difficult to deal with the limits of min or max

Smoothed margin

- The trick is to consider the smoothed margin

$$\tilde{\gamma}(w) := \frac{\ell^{-1}(\mathscr{L}(w))}{\|w\|^L}$$

  - Then, we know that

$$\tilde{\gamma}(w) \leq \gamma(w) \leq \tilde{\gamma}(w) + \frac{\log n}{\|w\|^L}$$

- **Want-to-show.** Convergence of the smoothed margin to the max margin

$$\lim_{t \to \infty} \tilde{\gamma}(w_t) = \bar{\gamma}$$

Linear case

- We prove the claim for the linear model $f(x; w) = \langle w, x \rangle$ trained with the gradient flow.

Theorem 10.1.

Suppose that $\max_{i \in [n]} \|x_i\| \leq 1$. Then, we have

$$\tilde{\gamma}(w(t)) \geq \bar{\gamma} - \frac{\log n}{\log t + \log(2n\bar{\gamma}^2) - 2\log\log(2tne\bar{\gamma}^2)}$$

- **Remark.** By noticing that $\tilde{\gamma}(w(t)) \leq \gamma(w(t)) \leq \bar{\gamma}$, we have a sandwich bound:

$$\bar{\gamma} - \frac{C}{\log t} \leq \gamma(w(t)) \leq \bar{\gamma}$$

  - Thus the limit converges

Proof sketch

- We track the numerator and the denominator of the smoothed margin, separately.

$$\tilde{\gamma}(w(t)) = \frac{\ell^{-1}(\mathscr{L}(w(t)))}{\|w(t)\|^1} =: \frac{u(t)}{v(t)}$$

- To bound $\tilde{\gamma}$ from below, we need:
  - A lower bound on

$$u(t) = -\log(\mathscr{L}(w(t)))$$

  - An upper bound on

$$v(t) = \|w(t)\|$$

Proof sketch

LB on $u(t) = -\log(\mathscr{L}(w(t)))$

- We have that

$$u(t) = u(0) + \int_0^t \dot{u}(s)\,\mathrm{d}s$$

- To analyze the first term, use the fact that we start GF at 0

$$u(0) = \ell^{-1}\left(\sum_{i=1}^{n} \ell(y_i \cdot \langle w, x_i \rangle)\right) = \ell^{-1}(n) = -\log(n)$$

- To analyze the second term, we begin by noticing that

$$\dot{u}(s) = \left\langle \frac{-\nabla \mathscr{L}(w(s))}{\mathscr{L}(w(s))}, \dot{w}(s) \right\rangle = \frac{\|\dot{w}(s)\|^2}{\mathscr{L}(w(s))}$$

Proof sketch

- Looking at $\|\dot{w}(s)\|$, we have

$$\begin{aligned}
\|\dot{w}(s)\| \geq \langle \dot{w}(s), \bar{u} \rangle &= \left\langle -\sum_{i=1}^{n} x_i y_i \ell'(m_i(w(s))), \bar{u} \right\rangle \\
&= \sum_{i=1}^{n} \ell(m_i(w(s))) \cdot \langle y_i x_i, \bar{u} \rangle \\
&\geq \bar{\gamma} \cdot \sum_{i=1}^{n} \ell(m_i(w(s))) \\
&= \bar{\gamma} \cdot \mathscr{L}(w(s))
\end{aligned}$$

Proof sketch

- Plugging in, we get

$$
\begin{aligned}
u(t) &= u(0) + \int_0^t \dot{u}(s)\,\mathrm{d}s \\
&\geq -\log(n) + \int_0^t \frac{\|\dot{w}(s)\|^2}{\mathscr{L}(w(s))}\,\mathrm{d}s \\
&\geq -\log(n) + \bar{\gamma} \cdot \int_0^t \|\dot{w}(s)\|\,\mathrm{d}s \\
&\geq -\log(n) + \bar{\gamma} \cdot \|\dot{w}(t)\|
\end{aligned}
$$

Proof sketch

UB on $v(t) = \|w(t)\|$

- We have that

$$v(t) = \|w(t)\| \leq \int_0^t \|\dot{w}(s)\| \, \mathrm{d}s$$

  - where

$$\begin{aligned} \|\dot{w}(s)\| &= \left\| \sum_{i=1}^n x_i y_i \ell'(m_i(w(s))) \right\| \\ &\leq \sum_{i=1}^n \ell(m_i(w(s))) \\ &\leq \mathscr{L}(w(s)) \end{aligned}$$

- By the risk convergence (Lemma 10.1), we have UB of form $\log s/s$ on the RHS
  - Integrate, and get what we want

Next up

- Generalization