13. Linearization - 2

This slide

- Utilize classical optimization tools, but for neural nets

- **Idea.** Consider the linearized model, i.e., NTK regime
    - Happens near initialization
    - Happens for overparameterized model

    - Today. Happens for scaled-up initial models $f \mapsto \alpha \cdot f$
        - Mainly follow the proof of Chizat and Bach (2019)
            - "On Lazy Training in Differentiable Programming" NeurIPS 2019

Recall

- Neural nets near initialization are almost linear:

$$f_0(\mathbf{x};\mathbf{w}) = f(\mathbf{x};\mathbf{w}_0) + \langle \partial_{\mathbf{w}} f(\mathbf{x};\mathbf{w}_0), \mathbf{w}-\mathbf{w}_0\rangle$$

  - For smooth activations, we had:

$$f(\mathbf{x};\mathbf{w}) - f_0(\mathbf{x};\mathbf{w}) \le C \cdot \|\mathbf{w}-\mathbf{w}_0\|_F^2 \,/m^{1/2}$$

  - For ReLU nets, we had:

$$f(\mathbf{x};\mathbf{w}) - f_0(\mathbf{x};\mathbf{w}) \le C \cdot \|\mathbf{w}-\mathbf{w}_0\|_F^{4/3} \,/m^{1/3}$$

- The linearized models are universal approximators

Setup

- **Question.** When we run GD, do we stay close to the initialization?

- **Notation.** We bake the training set into the predictor

$$f(\mathbf{w}) = [f(\mathbf{x}_1; \mathbf{w}), f(\mathbf{x}_2; \mathbf{w}), \cdots, f(\mathbf{x}_n; \mathbf{w})]^\top \in \mathbb{R}^n$$

- **Problem.** The squared loss regression, with a scale factor α

$$\hat{R}(\alpha \cdot f(\mathbf{w})) := \frac{1}{2}\|y - \alpha \cdot f(\mathbf{w})\|^2$$

$$\hat{R}_0 = \hat{R}(\alpha \cdot f(\mathbf{w}(0)))$$

Setup

- **Optimizer.** We consider the gradient flow $\mathbf{w}(t)$

$$\begin{aligned}\dot{\mathbf{w}}(t) &:= -\nabla_{\mathbf{w}}\hat{R}\big(\alpha \cdot f(\mathbf{w}(t))\big) \\ &= -\alpha J_t^\top \nabla \hat{R}\big(\alpha \cdot f(\mathbf{w}(t))\big)\end{aligned}$$

  - Here, J_t denotes the Jacobian

$$J_t = \begin{bmatrix} \nabla f(\mathbf{x}_1; \mathbf{w}(t))^\top \\ \cdots \\ \nabla f(\mathbf{x}_n; \mathbf{w}(t))^\top \end{bmatrix} \in \mathbb{R}^{n \times p}$$

Setup

- We denote the linear approximation of $\mathbf{w}(t)$ by $\mathbf{u}(t)$

$$f_0(\mathbf{u}) := f(\mathbf{w}(0)) + J_0(\mathbf{u} - \mathbf{w}(0))$$

- The trajectory of $\mathbf{u}(t)$ is given by

$$\begin{aligned}\dot{\mathbf{u}}(t) &= -\nabla_{\mathbf{u}} \hat{R}\left(\alpha \cdot f_0(\mathbf{u}(t))\right) \\ &= -\alpha \cdot J_0^\top \nabla \hat{R}\left(\alpha \cdot f_0(\mathbf{u}(t))\right)\end{aligned}$$

- **Goal.** Show that, for nice α, we have
  - Both $\mathbf{w}(t)$ and $\mathbf{u}(t)$ stays close to $\mathbf{w}(0) = \mathbf{u}(0)$
    - i.e., safe to use guarantees for the linearization
  - Both $f(\mathbf{w}(t))$ and $f_0(\mathbf{u}(t))$ achieves small risks

Assumptions

- We impose some assumptions on the Jacobian J_t
  - $\text{rank}(J_0) = n$
    - exact solution exists for the f_0
  - $\sigma_{\min} := \sigma_{\min}(J_0) = \sqrt{\lambda_{\min}(J_0 J_0^\top)} > 0$
  - $\sigma_{\max} > 0$
  - $\|J_w - J_v\| \leq \beta \|w - v\|$

Main result

Theorem 8.1.

Assume that we have

$$\alpha \geq \beta \sqrt{1152 \cdot \sigma_{\max}^2 \hat{R}_0} \Big/ \sigma_{\min}^3$$

Then, we have:

- $\hat{R}\big(\alpha \cdot f(\mathbf{w}(t))\big) \leq \hat{R}_0 \cdot \exp\left(-t\alpha^2\sigma_{\min}^2/2\right)$
- $\hat{R}\big(\alpha \cdot f_0(\mathbf{u}(t))\big) \leq \hat{R}_0 \cdot \exp\left(-t\alpha^2\sigma_{\min}^2/2\right)$

Also, we have

- $\|\mathbf{w}(t) - \mathbf{w}(0)\| \leq \sqrt{72 \cdot \sigma_{\max}^2 \cdot \hat{R}_0} \Big/ \alpha \cdot \sigma_{\min}^2$
- $\|\mathbf{u}(t) - \mathbf{u}(0)\| \leq \sqrt{72 \cdot \sigma_{\max}^2 \cdot \hat{R}_0} \Big/ \alpha \cdot \sigma_{\min}^2$

- Exponential convergence of risk & parameter stays within a constant range

Main result

- The theorem depends on a lot of quantities
    - Smoothness constant β
    - Singular values $\sigma_{\min}, \sigma_{\max}$
    - Initial risk $\hat{R}_0$

- Before proving, let's get used to these quantities

Case study: Shallow neural net

- Consider a shallow neural net

$$f(\mathbf{x};\mathbf{w}) = \sum_j s_j \sigma(\mathbf{w}_j^\top \mathbf{x})$$

  - Here, s_j are non-trainable binary weights, i.e., $s_j \in \{-1, +1\}$

- **Jacobian.** Can be written as:

$$J_w = \begin{bmatrix} s_1\sigma'(\mathbf{w}_1^\top \mathbf{x}_1)\mathbf{x}_1^\top, & \cdots & s_m\sigma'(\mathbf{w}_m^\top \mathbf{x}_1)\mathbf{x}_1^\top \\ & \cdots & \\ s_1\sigma'(\mathbf{w}_1^\top \mathbf{x}_n)\mathbf{x}_n^\top, & \cdots & s_m\sigma'(\mathbf{w}_m^\top \mathbf{x}_n)\mathbf{x}_n^\top \end{bmatrix}$$

Case study: Shallow neural net

- **Smoothness.** If the activation function is β_0-smooth, then we have

$$
\begin{aligned}
\|J_w - J_v\|^2 &= \sum_{i=1}^{n} \sum_{j=1}^{m} s_j^2 \|\mathbf{x}_i\|^2 (\sigma'(\mathbf{w}_j^\top \mathbf{x}_i) - \sigma'(\mathbf{v}_j^\top \mathbf{x}_i))^2 \\
&= \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 \Big(\sum_{j=1}^{m} (\sigma'(\mathbf{w}_j^\top \mathbf{x}_i) - \sigma'(\mathbf{v}_j^\top \mathbf{x}_i))^2 \Big) \\
&\le \beta_0^2 \sum_{i=1}^{n} \|\mathbf{x}_i\|^2 \Big(\sum_{j=1}^{m} \|\mathbf{w}_j - \mathbf{v}_j\|^2 \|\mathbf{x}_i\|^2 \Big) \\
&\le \beta_0^2 \cdot \Big(\sum_{i=1}^{n} \|\mathbf{x}_i\|^4 \Big) \cdot \|\mathbf{w} - \mathbf{v}\|^2
\end{aligned}
$$

Case study: Shallow neural net

- **Singular values.** Consider the entries of the matrix

$$(J_0 J_0^\top)_{i,j} = \nabla f(\mathbf{x}_i; \mathbf{w}(0))^\top \nabla f(\mathbf{x}_j; \mathbf{w}(0))$$

  - At initialization, we may assume that each vector of $\mathbf{w}(0)$ is an i.i.d. copy of some random $\mathbf{v}$

  - Then, we have

$$\begin{aligned} \mathbb{E}(J_0 J_0^\top)_{i,j} &= \mathbb{E}\left[\sum_k s_k^2 \cdot \sigma'(\mathbf{w}_k(0)^\top \mathbf{x}_i) \cdot \sigma'(\mathbf{w}_k(0)^\top \mathbf{x}_j) \cdot \mathbf{x}_i^\top \mathbf{x}_j\right] \\ &= m \cdot \mathbb{E}\left[\sigma'(v^\top \mathbf{x}_i) \cdot \sigma'(v^\top \mathbf{x}_j) \cdot \mathbf{x}_i^\top \mathbf{x}_j\right] \end{aligned}$$

  - Thus, it is natural to expect that

$$\sigma_{\max}, \sigma_{\min} \propto \sqrt{m}$$

Case study: Shallow neural net

- **Initial risk.** Suppose that we draw

$$s_i \sim \mathrm{Unif}(\{+1, -1\}), \qquad \mathbf{w}_i \sim P$$

  - Then, we have

$$\begin{aligned}
\mathbb{E}\hat{R}_0 &= \mathbb{E}\Big[\sum_{i=1}^{n} \frac{1}{2}\big(y_i - \alpha \cdot f(\mathbf{x}_i; \mathbf{w}(0))\big)^2\Big] \\
&= \frac{1}{2}\sum_{i=1}^{n} \|y_i\|^2 + \frac{\alpha^2}{2}\sum_{i=1}^{n} \mathbb{E}\|f(\mathbf{x}_i; \mathbf{w}(0))\|^2 \\
&= \Theta(\alpha^2 mn)
\end{aligned}$$

- Combining all these, we see that the assumption $\alpha \geq \beta\sqrt{1152 \cdot \sigma_{\max}^2 \hat{R}_0} \Big/ \sigma_{\min}^3$ actually means that the model is sufficiently wide, comparing with the number of data.

Proof plan

- Choose some radius B
    - Consider a ball

$$\mathscr{B} = \{\mathbf{v} \mid \|\mathbf{v} - \mathbf{w}(0)\| \leq B\}$$

    - Choose

$$T := \inf\{t \geq 0 \;:\; \|\mathbf{w}(t) - \mathbf{w}(0)\| > B\}$$

- For any $t \in [0,T]$:
    - If $J_t J_t^\top$ is positive-definite, risk decreases rapidly (Lemma 8.1.)
    - Rapid risk decrease —> Cannot travel far (Lemma 8.2.)
    - These holds for $\mathbf{u}(t)$, as $J_0 J_0^\top$ is positive-definite
        - For $\mathbf{w}(t)$, additional work is needed (Lemma 8.3.; not discussed today)

Evolution of predictions

- Let us look at how predictions evolve

- **Original.** Difficult to track J_t

$$\frac{\mathrm{d}}{\mathrm{d}t}\alpha f(\mathbf{w}(t)) = \alpha J_t \dot{\mathbf{w}}(t) = -\alpha^2 J_t J_t^\top \nabla \hat{R}(\alpha f(\mathbf{w}(t)))$$
$$= -\alpha^2 J_t J_t^\top (\alpha f(\mathbf{w}(t)) - y)$$

- **Linearized.** Easier to track — becomes convex quadratic

$$\frac{\mathrm{d}}{\mathrm{d}t}\alpha f_0(\mathbf{u}(t)) = \alpha J_0 \dot{\mathbf{u}}(t)$$
$$= -\alpha^2 J_0 J_0^\top \nabla \hat{R}(\alpha f_0(\mathbf{u}(t)))$$
$$= -\alpha^2 J_0 J_0^\top (\alpha f_0(\mathbf{u}(t)) - y)$$

- For original to converge, we may need a uniform control over $J_t J_t^\top$

Rapid decay of risk

Lemma 8.1.

Suppose that we have some GF trajectory $\mathbf{z}(t)$ with

$$\dot{\mathbf{z}}(t) = -\,Q(t)\,\nabla\hat{R}(\mathbf{z}(t)).$$

Define the minimum eigenvalue

$$\lambda := \inf_{t\in[0,\tau]} \lambda_{\min}\big(Q(t)\big) > 0$$

Then, for any $t \in [0,\tau]$, we have

$$\hat{R}(\mathbf{z}(t)) \le \hat{R}(\mathbf{z}(0)) \cdot \exp(-2\lambda t)$$

- **Interpretation.** Uniform lower bound means exponential convergence

Rapid decay of risk

Lemma 8.1.

Suppose that we have some GF trajectory $\mathbf{z}(t)$ with

$$\dot{\mathbf{z}}(t) = -Q(t)\nabla\hat{R}(\mathbf{z}(t)).$$

Define the minimum eigenvalue

$$\lambda := \inf_{t\in[0,\tau]} \lambda_{\min}\big(Q(t)\big) > 0$$

Then, for any $t \in [0,\tau]$, we have

$$\hat{R}(\mathbf{z}(t)) \le \hat{R}(\mathbf{z}(0)) \cdot \exp(-2\lambda t)$$

- **Interpretation.** Uniform lower bound means exponential convergence

Proof sketch

- Proceed as:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|\mathbf{z}(t)-y\|^2 &= \langle -Q(t)(\mathbf{z}(t)-y), \mathbf{z}(t)-y\rangle \\
&\leq -\lambda_{\min}\big(Q(t)\big)\cdot\|\mathbf{z}(t)-y\|^2 \\
&\leq -2\lambda\cdot\left(\frac{1}{2}\|\mathbf{z}(t)-y\|^2\right)
\end{aligned}
$$

- Then, apply the Grönwall's inequality

Trajectory stays within the ball

Lemma 8.2.

Suppose that

$$\dot{\mathbf{v}}(t) = -\,S(t)^\top \nabla \hat{R}(g(\mathbf{v}(t))).$$

where we know that

$$\lambda_i\left(S_t S_t^\top\right) \in [\lambda, \lambda_1] \qquad \forall t \in [0,\tau]$$

Then, for any $t \in [0,\tau]$, we have

$$\|\mathbf{v}(t) - \mathbf{v}(0)\| \leq \frac{\sqrt{\lambda_1}}{\lambda}\|g(\mathbf{v}(0)) - y\| \leq \frac{\sqrt{2\lambda_1 \hat{R}(g(\mathbf{v}(0))}}{\lambda}$$

- **Interpretation.** If eigenvalues admit uniform upper and lower bounds, the trajectory stays within some ball

Proof sketch

- Proceed as:

$$\begin{aligned}
\|\mathbf{v}(t) - \mathbf{v}(0)\| &= \left\| \int_0^t \dot{\mathbf{v}}(s)\,\mathrm{d}s \right\| \leq \int_0^t \|\dot{\mathbf{v}}(s)\|\,\mathrm{d}s \\
&= \int_0^t \|S_t^\top \nabla \hat{R}(g(\mathbf{v}(s)))\|\,\mathrm{d}s \\
&\leq \sqrt{\lambda_1} \int_0^t \|g(\mathbf{v}(s)) - y\|\,\mathrm{d}s \\
&\leq \sqrt{\lambda_1} \|g(\mathbf{v}(0)) - y\| \int_0^t \exp(-s\lambda)\,\mathrm{d}s \\
&\leq \frac{\sqrt{\lambda_1}}{\lambda} \|g(\mathbf{v}(0)) - y\|
\end{aligned}$$

Eigenvalue analysis

- For $\mathbf{u}(t)$, we can evaluate the eigenvalues of $J_0 J_0^\top$ fairly well
    - Simply use $\sigma_{\min}, \sigma_{\max}$
- For $\mathbf{w}(t)$, we need some additional work
    - See Lemma 8.3. in the textbook