10/30/23

Lecture 15: Linearized networks

NTK analysis :

in fact we'll prove a generic result that doesn't even need the assumption that the student network is a one-hidden-layer MLP.

Consider a dataset $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ and a student network $f_\theta : \mathbb{R}^d \to \mathbb{R}$, $\theta \in \mathbb{R}^p$, initialized to some θ_0

We'll use shorthand $f_\theta(x) - y$ to denote

$$\left(f_\theta(x_1) - y_1, \; \dots, \; f_\theta(x_n) - y_n \right).$$

Define: scaling param: $\alpha > 0$

empirical loss: $\hat{L}(g) \triangleq \frac{1}{2} \| g(x) - y \|_2^2$

$$\hat{L}_0 \triangleq \hat{L}(\gamma f_{\theta_0})$$

gradient flow :

$$d\theta_t \triangleq - \nabla_\theta \hat{L}(\gamma f_{\theta_t})\, dt$$

$$= -\gamma J_t^T \nabla \hat{L}(\gamma f_{\theta_t})\, dt, \quad \text{where}$$

Jacobian : $J_t \triangleq J_{\theta_t} \triangleq \begin{pmatrix} - \nabla_\theta f_{\theta_t}(x_1) - \\ \vdots \\ - \nabla_\theta f_{\theta_t}(x_n) - \end{pmatrix} \in \mathbb{R}^{n \times p}$

Will compare to linearized network / dynamics:

$$f^{lin}_{\theta}(x) = f_{\theta_0}(x) + J_0 \cdot (\theta - \theta_0)$$

$$d\tilde{\theta}_t \triangleq -\nabla_\theta \hat{L}(\gamma f^{lin}_{\tilde{\theta}_t})$$

$$= -\gamma \boxed{J_0^T} \nabla \hat{L}(\gamma f^{lin}_{\tilde{\theta}_t})$$

Jacobian does not change for linearized network

Will assume that

1). J_θ is Lipschitz in θ, i.e.

$$\| J_\theta - J_{\theta'} \|_{op} \leq \beta \| \theta - \theta' \|_2$$

2) $J_0 = J_{\theta_0}$ is full rank (bounds will depend on σ_{min}, σ_{max} of J_0)

Linearized dynamics very easy to analyze:

Lemma 1: If $Q(t) \succeq \lambda \cdot Id_n \ \forall t$, then for (g_t) given by

$$dg_t = -Q(t) \nabla \hat{L}(g_t)\, dt,$$

we have

$$\hat{L}(g_t) \leq \hat{L}(g_0) \cdot \exp(-2\lambda t).$$

pf: $\frac{d}{dt}\hat{L}(g_t) = \langle -Q(t)(g_t(x)-y), g_t(x)-y \rangle$

$$\leq -\lambda \|g_t(x)-y\|_2^2$$

$$= -2\lambda \cdot \hat{L}(g_t),$$

So integrating this (i.e. using Grönwall's inequality) completes the proof. $\square$

Can apply this to $Q(t) = J_t J_t^T$ and $g_t = \gamma f^{lin}_{\hat{\theta}_t}$.

Then b/c

$$d\hat{\theta}_t = -\gamma J_0^T \nabla_\theta \hat{L}(\gamma f^{lin}_{\hat{\theta}_t})\, dt,$$

by chain rule,

$$\frac{d}{dt}\left(\gamma f^{lin}_{\hat{\theta}_t}\right) = \gamma \underbrace{\nabla_\theta f^{lin}_\theta \Big|_{\theta=\hat{\theta}_t}}_{J_0} \cdot \frac{d\hat{\theta}_t}{dt}$$

$$= -\gamma^2 \underbrace{J_0 J_0^T}_{Q(t) \succeq \sigma^2_{min}(J_0)\cdot I_d} \nabla_\theta \hat{L}(\gamma f^{lin}_{\hat{\theta}_t}),$$

so by Lemma, $\hat{L}(\gamma f^{lin}_{\hat{\theta}_t}) \leq \exp(-2\gamma^2 t \sigma^2_{min}(J_0))$

So (unsurprisingly), training loss for linearized network drops exponentially quickly.

Can also show that, relative to drop in loss, movement of parameters is negligible:

<u>Lemma 2</u>: Suppose process $(\hat{\theta}_t)$ satisfies

$$d\hat{\theta}_t = -S(t)^T \nabla \hat{L}(g_{\hat{\theta}_t})$$

for some network g, and $\underline{\lambda}\cdot Id \preceq S(t)S(t)^T \preceq \bar{\lambda}\cdot Id \;\; \forall t$. Then

$$\|\hat{\theta}_t - \hat{\theta}_0\| \le \frac{\sqrt{\bar{\lambda}}}{\underline{\lambda}} \|g_{\hat{\theta}_0}(x) - y\|$$

<u>Pf</u>: $$\|\hat{\theta}_t - \hat{\theta}_0\| = \left\| \int_0^t \left(-S(s)^T \nabla \hat{L}(g_{\hat{\theta}_s})\right) ds \right\|$$

$$\le \int_0^t \underbrace{\|S(s)\|_{op}}_{\le \sqrt{\bar{\lambda}}} \cdot \underbrace{\|\nabla \hat{L}(g_{\hat{\theta}_s})\|}_{g_{\hat{\theta}_s}(x) - y} ds$$

$$\le \sqrt{\bar{\lambda}} \cdot \int_0^t \underbrace{\|g_{\hat{\theta}_s}(x) - y\|}_{\le \exp(-\lambda s)\cdot \|g_{\hat{\theta}_0}(x) - y\| \text{ (by prev. lemma)}} ds$$

$$\le \sqrt{\bar{\lambda}} \cdot \|g_{\hat{\theta}_0}(x) - y\| \cdot \underbrace{\int_0^s \exp(-\lambda s)\, ds}_{1/\lambda} . \quad \square$$

Applying this to $\hat{\theta}_t = \tilde{\theta}_t$, $g_\theta = f^{lin}_\theta$, $S(t) = \gamma J_0$,

$$\Rightarrow \|\tilde{\theta}_t - \theta_0\| \leq \frac{\sqrt{\sigma^2_{max}(J_0)}}{\sigma^2_{min}(J_0)} \cdot \|f_{\theta_0}(x) - y\|$$

$$\leq \frac{\sqrt{2\gamma^2\sigma^2_{max}(J_0)}}{\gamma^2\sigma^2_{min}(J_0)} \cdot \sqrt{\hat{L}_0}$$

$$\lesssim \frac{\sigma_{max}(J_0)}{\gamma\,\sigma^2_{min}(J_0)} \cdot \sqrt{\hat{L}_0}$$

Remains to show can apply Lemmas 1+2 to (θ_t). Complication is that J_t is changing over time. We will show it does not change that much, provided γ sufficiently large and θ_t remains close to initialization.

<u>Lemma 3</u>: If $\|\theta - \theta_0\| \leq \frac{\sigma_{min}(J_0)}{2\beta} \triangleq B$, then

$$\frac{\sigma_{min}(J_0)}{2} \cdot I_d \preceq J_\theta \preceq \frac{3\,\sigma_{max}(J_0)}{2} \cdot I_d$$

Pf:

$$
\begin{aligned}
\|J_\theta\|_{op} &\le \|J_0\|_{op} + \|J_\theta - J_0\|_{op} \\
&\le \sigma_{max}(J_0) + \underbrace{\beta\|\theta - \theta_0\|}_{\le \frac{\sigma_{min}(J_0)}{2} \le \frac{\sigma_{max}(J_0)}{2}} \\
&\le \frac{3\sigma_{max}(J_0)}{2}.
\end{aligned}
$$

for lower bound, for any $\|v\| = 1$,

$$
\begin{aligned}
\|J_\theta v\| &\ge \|J_0 v\| - \|(J_\theta - J_0)v\| \\
&\ge \|J_0 v\| - \beta\|\theta - \theta_0\| \ge \frac{\sigma_{min}(J_0)}{2}. \ \square
\end{aligned}
$$

So we can safely apply Lemma 1+2 to get

$$
\hat{L}(\gamma f_{\theta_t}) \le \exp\left(-\tfrac{1}{2}\gamma^2 \sigma_{min}^2(J_0)\, t\right)
$$

$$
\|\theta_t - \theta_0\| \lesssim \frac{\sigma_{max}}{\gamma \sigma_{min}^2} \cdot \sqrt{\hat{L}_0} \qquad (*)
$$

as long as $\|\theta_s - \theta_0\| \le B \quad \forall \ s \in [0, t]$

note that bound in $(*)$ $\ll B$ as long as

$$
\gamma \gg \frac{\beta \sigma_{max}}{\sigma_{min}^3} \sqrt{\hat{L}_0}.
$$

We conclude

Thm: Linearized network $f^{lin}_{\hat{\theta}_t}$ and true network f_{θ_t} stay $\frac{\sigma_{max}}{\gamma \sigma^2_{min}} \sqrt{\hat{L}_0}$ -close for all $t \geq 0$, and training loss for f_{θ_t} drops exponentially quickly.

Example: Consider $\gamma f_\theta = \gamma \sum_{i=1}^{N} a_i \sigma(\langle w_i, x \rangle)$. For simplicity, suppose a_i's are random $\{\pm 1\}$'s that are not subsequently trained, so $\theta = \{w_i\}_{i=1}^N$.

$\boxed{\beta}$
$$J_\theta = \begin{pmatrix} x_1^T \cdot \{a_i \sigma'(\langle w_i, x \rangle)\}_i \\ \vdots \\ x_n^T \cdot \{a_i \sigma'(\langle w_i, x \rangle)\}_i \end{pmatrix}$$

So $\|J_\theta - J_{\theta'}\|_{op}^2$

$$\leq \sum_{i,j} \|x_i\|_2^2 \cdot \underbrace{\left(\sigma'(\langle w_j, x \rangle) - \sigma'(\langle w'_j, x \rangle)\right)^2}_{\lesssim \|w_j - w'_j\|^2}$$

$$= \|X\|_F^4 \cdot \|\theta - \theta'\|_2^2$$

So can take $\beta \asymp \|X\|_F^2 \asymp \underline{nd}$ (e.g. if $x \sim S^{d-1} \cdot \sqrt{d}$)

$\boxed{\hat{L}_0}$: Can initialize in such a way that $f_{\theta_0}(x)$ is dominated by y. So

$$\hat{L}_0 \asymp \|y\|^2 \asymp n$$

$\boxed{\sigma_{min}(J_0), \sigma_{max}(J_0)}$: entries of J_0 are $O(1)$, so because $Nd \gg n$, singular values are of order $\sqrt{Nd}$

Putting everything together, $\gamma \gg \frac{\beta \sigma_{min}(J_0)}{\sigma^2_{max}(J_0)} \sqrt{\hat{L}_0}$

yields

$$\gamma \gg \frac{nd \cdot \sqrt{Nd}}{(\sqrt{Nd})^3} \cdot \sqrt{n}$$

$$= \frac{n^{3/2}}{N},$$

so provided we are in this regime, gradient flow well-approx'd by linearized dynamics.